CPATH τεχνη Evaluation: Summer ’09

Robert J. Schalkoff
rjschal@clemson.edu
Clemson University
Department of Electrical and Computer Engineering

rev. July 22, 2009
The SRI ("super evaluator") effort has been substantially redefined:

- SRI payroll reduced
- 'All-inclusive' web experiment less ambitious

Evaluators meeting in October 2009.

I have been asked to document and disseminate some results, possibly in August 2009.
Are there quantitative (incremental?) benefits from the τεχνη approach? (student motivation, learning outcomes, instructor motivation)
The proposed Level 2 evaluation design is a Nonequivalent Group with Pre/Post Test Design\(^1\), typically diagrammed as:

\[
\begin{array}{ccc}
\text{NR} & O_1 & X & O_2 \\
\text{NR} & O_1 & O_2 \\
\end{array}
\]

where \(X\) represents \(\tau \epsilon \chi \nu \eta\) (the treatment).

\(^1\text{For the trials in Fall 2008, it was be a Post-test only.}\)
Initial cohorts for the quasi-experimental design were proposed in the Fall of 2008 and fall into 4 main course categories.

**Table 1: Quasi-experimental Design Cohorts and Course Coordinators, NEGD and Level-3 (Proposed Fall 2008).**

<table>
<thead>
<tr>
<th>Ref</th>
<th>School</th>
<th>Course</th>
<th>Coordinator</th>
<th>Pedagogy</th>
<th>Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UNC-W</td>
<td>CSC-112</td>
<td>Narayan</td>
<td>τεχνη</td>
<td>F08</td>
</tr>
<tr>
<td>2</td>
<td>CLM</td>
<td>CPSC 111</td>
<td>Duchowski</td>
<td>conventional</td>
<td>F08</td>
</tr>
<tr>
<td>3</td>
<td>WCU</td>
<td>CS 140</td>
<td>Dalton</td>
<td>conventional</td>
<td>F08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CLM</td>
<td>CPSC 101</td>
<td>Westall</td>
<td>τεχνη</td>
<td>F08</td>
</tr>
<tr>
<td>5</td>
<td>WCU</td>
<td>CS 150</td>
<td>Holliday</td>
<td>conventional</td>
<td>F08</td>
</tr>
<tr>
<td>6</td>
<td>UNC-W</td>
<td>CSC 121</td>
<td>Narayan</td>
<td>conventional</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CLM</td>
<td>CPSC 212</td>
<td>Geist</td>
<td>τεχνη</td>
<td>F08</td>
</tr>
<tr>
<td>8</td>
<td>WCU</td>
<td>CS 351</td>
<td>Dalton</td>
<td>conventional</td>
<td>Sp09</td>
</tr>
<tr>
<td>9</td>
<td>UNC-W</td>
<td>CSC 332</td>
<td>Narayan</td>
<td>conventional</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CLM</td>
<td>CPSC 462</td>
<td>Pargas/Wang</td>
<td>τεχνη</td>
<td>F08/Sp09</td>
</tr>
<tr>
<td>11</td>
<td>WCU</td>
<td>CS 453</td>
<td>Holliday</td>
<td>τεχνη</td>
<td>Sp09</td>
</tr>
<tr>
<td>12</td>
<td>UNC-W</td>
<td>CSC 455</td>
<td>Narayan</td>
<td>conventional</td>
<td></td>
</tr>
</tbody>
</table>
For the Fall 2008 semester, Groups QE1 and QE2 were evaluated with both Level-2 (sans pretest) and Level-3 instruments.

For the Spring 2009 semester, Group QE4 was evaluated with both Level-2 (including pretest) and Level-3 instruments. The PreTest consisted of 4 quantitative questions (9-12).

For the Spring 2009 semester, Groups QE1 and QE2 were evaluated with only Level-3 instruments. This is of secondary interest to NSF; analysis will occur later.

The future (Fall 2009 - Sp 2010): After lunch
Executive Summary of Results to Date

- **Level 2 Statistical Results**
  - QE1; Fall 2008; PostTest only: conventional more effective
  - QE2; Fall 2008; PostTest only: (depends upon which test used); \( \tau \epsilon \chi \nu \eta \) more effective
  - QE4; Sp 2009; Pre and PostTest: inconclusive (possibly re-instrument and need more data)

- **Level 3 Preliminary Results (All groups):** Surveys seem to indicate some perception that \( \tau \epsilon \chi \nu \eta \) experience generates more enthusiasm and is more 'real'.
Level 3 Data Analysis Strategy

- **Post Test only**: $t$, Welsh and relatives
- **Pre/Post Test**: ANACOVA/OLS
There was no pretest (PostTest only).
In these Figures, the y-axis shows an average of the embedded problem grades. The integer grading scale used ranged from 0 (no solution or clueless) to 4 (perfect score).
QE1 Analysis

Notes:
QE1-1 is TEXNH
QE1-2 and QE1-3 are conventional (lumped)

IMPLEMENTATION OF HYPOTHESIS TESTING
avg (mu_i) std number
texnh 1.89 0.7004 16 n1
conventional 2.25 0.94 21 n2

STUDENT’S T assumes equal but unknown variances
H0: mu_1=mu_2 H1: mu_1<mu_2 Alpha=0.05 (one tailed)
nu 35
Sp^2 0.7106
Sp 0.8430
T -1.2874
t_alpha 1.6896 One-tailed; uses 2-tailed formula
test T < -t_alpha FALSE
Result: reject H_0, accept H_1

Welch’s t-test
nu 34.9966 num 0.005235
T’-1.3388 den 0.000150
t_alpha 1.6909 ratio 34.996619
here T’ > t_alpha; reject H0; accept H1
In these Figures, the y-axis shows an average of the embedded problem grades. The integer grading scale used ranged from 0 (no solution or clueless) to 4 (perfect score).

Figure 2: Fall 2008: QE2 Preliminary Composite Data, from Embedded Problems.
- Pool QE2-5 and QE2-6 cohorts into single 'conventional' cohort

- Resulting statistics:

<table>
<thead>
<tr>
<th></th>
<th>$\mu_i^s$</th>
<th>$\sigma_i^s$ (std)</th>
<th>$n_i$ (number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\tau\varepsilon\chi\nu\eta$</td>
<td>2.81</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>conventional</td>
<td>2.38</td>
<td>1.16</td>
</tr>
</tbody>
</table>

Table 2: QE2 $\tau\varepsilon\chi\nu\eta$ and conventional (pooled QE2/QE3) statistics
Hypotheses:

\[ H_0 : \mu_1 = \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 = 0 \quad \text{no difference} \]

\[ H_1 : \mu_1 > \mu_2 \quad \text{or} \quad \mu_1 - \mu_2 > 0 \quad \text{significant difference} \]

Level of significance:

\[ \alpha = 0.05 \]
$T = \frac{\mu_{1s} - \mu_{2s}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

with $\sigma_1 = \sigma_2$ but unknown and $\nu = n_1 + n_2 - 2 = 63$ d.f.,

$$S_p^2 = \frac{(n_1 - 1)(\sigma_{1s})^2 + (n_2 - 1)(\sigma_{2s})^2}{\nu}$$

$S_p = 1.0336$

and the critical region is:

$T > t_\alpha$
For $\nu = n_1 + n_2 - 2 = 63$ d.f.,

$$t_\alpha = t_{0.05} = 1.669$$

(NIST table and spreadsheet; one-tailed)

Computed

$$T = 1.611$$

Conclusion:

$$T < t_\alpha \Rightarrow \text{accept } H_0$$
Probably more appropriate due to $\sigma^s_i$. Assume $\sigma_1 \neq \sigma_2$, but unknown.

$$T' = \frac{\mu^s_1 - \mu^s_2}{\sqrt{\frac{(\sigma^s_1)^2}{n_1} + \frac{(\sigma^s_2)^2}{n_2}}}$$

with

$$\nu = \frac{\left(\frac{(\sigma^s_1)^2}{n_1} + \frac{(\sigma^s_2)^2}{n_2}\right)^2}{\frac{(\sigma^s_1)^2}{n_1-1} + \frac{(\sigma^s_2)^2}{n_2-1}}$$

Here $T' = 1.79$, $\nu = 62$ and $t_{\alpha} = t_{0.05} = 1.67$

Conclusion: Since $T' > t_{0.05}$, reject $H_0$; accept $H_1$. 
The topics included satisfaction, effectiveness, opportunity to learn, assignments, engagement, learning. The survey format was 19 questions with an estimated completion time of 10 minutes.

7. I feel my classroom experience in this course generated enthusiasm for the subject.
   A) Very little   B) A little   C) Somewhat   D) A lot   E) A great deal

8. I feel my software development experience in this course generated enthusiasm for the subject.
   A) Very little   B) A little   C) Somewhat   D) A lot   E) A great deal

9. I feel the software development experience in this class used real-world examples.
   A) Very little   B) A little   C) Somewhat   D) A lot   E) A great deal
The y-axis indicates the per-unit response to a single question (#7).

Figure 4: Example of Distribution of Responses for τεχνη and Conventional Survey. Question #7 Shown. Leftmost (Darker) Column is τεχνη Data.
The y-axis represents an average of the weighted responses, where a 1 represents a response of 'Very little' and a 5 indicates a response of 'A great deal'. Leftmost (Darker) Column is Data.
Concern: group non-equivalence and/or instructor grading bias (probably requires adjusted pretest distribution).

<table>
<thead>
<tr>
<th>class</th>
<th>cohort</th>
<th>PRE</th>
<th>POST</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\tau \in \chi\nu\eta)</td>
<td>QE4-10</td>
<td>2.54</td>
<td>2.98</td>
</tr>
<tr>
<td>(\tau \in \chi\nu\eta)</td>
<td>QE4-11</td>
<td>2.04</td>
<td>3.00</td>
</tr>
<tr>
<td>conventional</td>
<td>QE4-12</td>
<td>2.68</td>
<td>3.63</td>
</tr>
</tbody>
</table>
Spring 2009 Evaluation: DBMS – Level 2 Scatter

Figure 7: Spring 2009: QE4 Data Scatter, from Embedded Problems.
HYPOTHESIS TESTING (pretest differences) Rev. 7-19-09

avg (mu_i) std number
texnh 2.447 0.8587 33 Aggregates QE4-10 and QE4-11 as 'texnh'
conventional 2.679 0.8794 14

STUDENT’S T (assumes equal but unknown variances)
H0: mu_1=mu_2 H1: mu_2>mu_1 Alpha=0.05 (one tailed)
nu 45
Sp^2 0.7477
Sp 0.8647
T 0.8397
t_alpha 1.6794 One-tailed accept H0

No.
IMPLEMENTATION OF HYPOTHESIS TESTING Rev. 6-09-09

FINAL SCORE avg (mu_i) std number

texnh 2.985 0.5363 34 Aggregates QE4-10 and QE4-11 as 'texnh'

conventional 3.625 0.3501 14

STUDENT’S T (assume equal but unknown variances)

H0: mu_1=mu_2 H1: mu_2>mu_1 Alpha=0.05 (one tailed)

nu 46
Sp^2 0.2410
Sp 0.4909
T 4.1070

t_alpha 1.6787 One-tailed Reject H0; accept H1
Normalize PostTest score with PreTest score average for cohort.

**HYPOTHESIS TESTING** (normout) Rev. 7-21-09

- normout avg (mu_i) std number
  - texnh 1.229 0.2651 33 Aggregates QE4-10 and QE4-11 as 'texnh'
  - conv. 1.353 0.1307 14 pretest avg for normalization

**STUDENT’S T** (assume equal but unknown variances)

- H0: mu_1=mu_2
- H1: mu_2>mu_1
- Alpha=0.05 (one tailed)
- nu 45
- Sp^2 0.0549
- Sp 0.2343
- T 1.6681
- t_alpha 1.6794 One-tailed accept H0
Each PreTest and corresponding PostTest score normalized (divided) by cohort average. Thus, cohorts all have 1.0 per unit PreTest average. change is difference of normalized Pre/PostTest scores.

HYPOTHESIS TESTING (normalized change) Rev. 7-21-09
Norm CHANGE avg (mu_i) std number
texnh 0.229 0.4126 33 Aggregates QE4-10 and QE4-11 as 'texnh'
conv. 0.353 0.3553 14

STUDENT’S T (assume equal but unknown variances)
H0: mu_1=mu_2 H1: mu_2>mu_1 Alpha=0.05 (one tailed)
nu 45
Sp^2 0.1575
Sp 0.3969
T 0.9847
t_alpha 1.6794 One-tailed accept H0
'Standard' regression using \textit{raw or unadjusted} pre-post scores and class:

\[
post = const + c_1 \times pre + c_2 \times class
\]

\(class = 1\) for texnh; \(class = 0\) for conventional

see results (next slide):

- +0.63 effect on PostTest from \textit{conventional}; conversely
- -0.63 penalty for \(\tau\varepsilon\chi\nu\eta\)
### Model 1: OLS estimates using the 47 observations 1–47

**Dependent variable:** post

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>3.51765</td>
<td>0.267155</td>
<td>13.1671</td>
<td>0.0000</td>
</tr>
<tr>
<td>pre</td>
<td>0.0400759</td>
<td>0.0863221</td>
<td>0.4643</td>
<td>0.6448</td>
</tr>
<tr>
<td>class</td>
<td>−0.630870</td>
<td>0.160956</td>
<td>−3.9195</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

- **Mean of dependent variable:** 3.17553
- **S.D. of dependent variable:** 0.573213
- **Sum of squared residuals:** 11.0321
- **Standard error of residuals (\( \hat{\sigma} \):** 0.500730
- **Unadjusted \( R^2 \):** 0.270089
- **Adjusted \( \bar{R}^2 \):** 0.236912
- **\( F(2, 44) \):** 8.14068
- **p-value for \( F() \):** 0.000981595
- **Log-likelihood:** -32.630
- **Akaike information criterion:** 71.2615
- **Schwarz Bayesian criterion:** 76.8119
- **Hannan–Quinn criterion:** 73.3501
Model 1:
OLS estimates using the 47 observations 1-47
Dependent variable: post

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STDERROR</th>
<th>T STAT</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1.30750</td>
<td>0.121988</td>
<td>10.718</td>
<td>&lt;0.00001 ***</td>
</tr>
<tr>
<td>pre</td>
<td>0.0453601</td>
<td>0.104435</td>
<td>0.434</td>
<td>0.66617</td>
</tr>
<tr>
<td>class</td>
<td>-0.124978</td>
<td>0.075235</td>
<td>-1.661</td>
<td>0.10379</td>
</tr>
</tbody>
</table>

Mean of dependent variable = 1.26511
Standard deviation of dep. var. = 0.238299
Sum of squared residuals = 2.44814
Standard error of residuals = 0.23588
Unadjusted R-squared = 0.06280
Adjusted R-squared = 0.02020
F-statistic (2, 44) = 1.47407 (p-value = 0.24)
Log-likelihood = 2.74813
Akaike information criterion (AIC) = 0.503744
Schwarz Bayesian criterion (BIC) = 6.05419
Hannan-Quinn criterion (HQC) = 2.59241
### QE4: Pre/PostTest Normalization (Model2–change)

Model 2:

OLS estimates using the 47 observations 1-47

Dependent variable: change

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STDERROR</th>
<th>T STAT</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>1.30629</td>
<td>0.122042</td>
<td>10.704</td>
<td>&lt;0.00001 ***</td>
</tr>
<tr>
<td>pre</td>
<td>-0.954150</td>
<td>0.104482</td>
<td>-9.132</td>
<td>&lt;0.00001 ***</td>
</tr>
<tr>
<td>class</td>
<td>-0.124567</td>
<td>0.0752684</td>
<td>-1.655</td>
<td>0.10505</td>
</tr>
</tbody>
</table>

Mean of dependent variable = 0.264681

Standard deviation of dep. var. = 0.396922

Sum of squared residuals = 2.45032

Standard error of residuals = 0.235985

Unadjusted R-squared = 0.66189

Adjusted R-squared = 0.64652

F-statistic (2, 44) = 43.0682 (p-value < 0.00001)

Log-likelihood = 2.72724

Akaike information criterion (AIC) = 0.545511

Schwarz Bayesian criterion (BIC) = 6.09595

Hannan-Quinn criterion (HQC) = 2.63418
Question 7
"my classroom experience in this course generated enthusiasm for the subject"
overall
3.26 Texnh
2.7 Conventional

Question 8
"my software development experience in this course generated enthusiasm for the subject"
overall
3.63 Texnh
2.5 Conventional

Question 9
"the software development experience in this class used real-world examples"
overall
4.23 Texnh
3.5 Conventional
Review: Summary of (Mixed) Results to Date

- **Level 2 Statistical Results**
  - QE1; Fall 2008; PostTest only: conventional more effective
  - QE2; Fall 2008; PostTest only: τεχνη more effective (using Welsh)
  - QE4; Sp 2009; Pre and PostTest: no difference or inconclusive (possibly re-instrument and need more data)

- **Level 3 Preliminary Results (All groups):** Surveys seem to indicate some perception that τεχνη experience generates more enthusiasm and is more 'real'.
The Fall 2008 and Spring 2009 experiments were illustrative and should improve our design and implementation for Fall 2009 and Spring 2010.

Noteworthy in the data reporting and survey implementation from the Fall 2008 were the pragmatics of collecting, selecting and distributing sets of common exam problems at the very end of a semester, as well as administering surveys. Spring 2009 only had minor glitches in grading and reporting (scale).

The Spring 2009 experiment may indicate potential selection bias or (more likely) instructor grading bias. We need to address this (later).
Minor Point: Issues With Spring 2009 Level-3 (survey)

- many students 'missed' back of pages
- multiple answers
- silly answers (handwritten)
Major Issues With QE4

- instrumentation threats:
  - pretest and/or posttest validity
  - model validity
  - grading variability?

- possibly nonequivalent groups??
Proposed Future Evaluation Efforts

Continued development and refinement of the evaluation instruments and procedures:

- Timetable for development of Level-2 instruments.
- Identification of cohorts and courses for Fall 2009 and beyond.
- Implementation (follow rubric; 0-4 vs. 0-5).
- Reporting.
- Keep evaluator in the loop.
Especially significant is uniform pre/post-test grading. We must maximize objectivity in the process (instructor-neutral).

- I suggest multiple choice.
- I suggest a larger number of questions.
- All cohorts should participate in exam design.
- Use common numbering system.