QUANTITATIVE EVIDENCE FOR THE USE OF SIMULATION AND RANDOMIZATION IN THE INTRODUCTORY STATISTICS COURSE

Nathan Tintle  
Associate Professor of Statistics  
Dordt College, Sioux Center, Iowa
Broader Context

- Randomization and simulation: What is it?
  - Simulation of null distributions
  - Bootstrapping
  - Permutation tests

- An incomplete recent history in the algebra-based introductory statistics course (Stat 101; AP Statistics equivalent)
  - Technological changes- implications on practice and teaching
  - Cobb (2005, 2007); renewed interest and catalyst
    - Deeper understanding of the logic and scope of inference
  - Modules and full length texts
Does it work?

- Anecdotal evidence; excitement; momentum; discussions; panels
- Quantitative evidence - Deeper understanding of logic and scope of inference?
  - Holcomb et al. ICOTS-8 (Slovenia; 2010)
  - Tintle, Joint Statistics Meeting, Panelist, Assessment results (2013)
Holcomb et al. 2010ab

- **Methods**
  - Modules introduced in the course
- **Key findings**
  - Not much improvement
- **Limitations**
  - Not a full curriculum implementation
  - Only one institution
Tintle et al. 2011

- **Methods**
  - One institution before and after switch
  - Full course redesign
  - Similar instructors before and after switch
  - Standardized assessment (CAOS; delMas et al. 2007)

- **Key findings**
  - Overall improved post-course performance
  - Areas with largest improvement in design and inference
  - ‘No harm’ in other areas

- **Limitations**
  - Conflation of design and pedagogy with ability to pinpoint reasons for improvement
Tintle et al. 2012

- **Methods**
  - One institution before and after switch
  - Full course redesign
  - Similar instructors before and after switch
  - Standardized assessment (CAOS; delMas et al. 2007)
  - Sub-sample measured 4 months post-course

- **Key findings**
  - Overall improved retention
  - Areas with largest improvement in retention were in design and inference
  - ‘No harm’ in other areas

- **Limitations**
  - Conflation of design and pedagogy with ability to pinpoint reasons for improved retention
Tinkle 2013

- **Methods**
  - New assessment instrument (modified CAOS ~30 questions)
  - Multiple institutions participating

- **Key findings**
  - Overall, similar results at other institutions

- **Limitations**
  - Not always a ‘before the change’ at the institution
  - Different institutions, pedagogies and uses of materials; large number of potential confounding variables
New results

- Another before and after story (Dordt College)
- Transferability – 2013/2014 results
- What about low performers?
Dordt’s before and after story

- **Methods**
  - *Traditional curriculum* (Moore 2010) - 94 students; spring 2011
  - New curriculum (ISI, 2011 version) – 155 students; fall 2011 and spring 2012
  - All students completed the 40-question CAOS test during the first week of the semester and again during the last week of the semester. Students were given course credit for completing the assessment test, but not for their performance, and the test was administered electronically outside of class.
  - Two instructors taught the course each semester, with one instructor the same each semester, and one different in spring 2011 than in fall 2011/spring 2012
Dordt’s before and after story

- Overall performance

Very similar to Tintle et al (2011) results at another institution. Approx. twice the gains using new curriculum as compared to traditional (11.6% vs. 5.6%; p<0.001)
### Dordt’s before and after story

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Cohort</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Diff.</th>
<th>Paired t-test p-value</th>
<th>Cohort p-value</th>
<th>95% CI for cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Collection and Design</strong></td>
<td>Random.</td>
<td>34.8%</td>
<td>53.1%</td>
<td>18.2%</td>
<td>&lt;0.001</td>
<td>0.54</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>34.9%</td>
<td>36.5%</td>
<td>1.6%</td>
<td>0.014</td>
<td>0.03</td>
<td>(-2.1%, -18.1%)</td>
</tr>
<tr>
<td><strong>Descript. Statistics</strong></td>
<td>Random.</td>
<td>55.1%</td>
<td>61.1%</td>
<td>6.0%</td>
<td>&lt;0.001</td>
<td>0.015</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>53.5%</td>
<td>69.6%</td>
<td>16.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Graphical Representations</strong></td>
<td>Random.</td>
<td>55.8%</td>
<td>64.4%</td>
<td>8.6%</td>
<td>&lt;0.001</td>
<td>0.03</td>
<td>(0.6%, 11.4%)</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>58.5%</td>
<td>60.9%</td>
<td>2.4%</td>
<td>0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Boxplots</strong></td>
<td>Random.</td>
<td>35.0%</td>
<td>41.6%</td>
<td>6.6%</td>
<td>0.010</td>
<td>0.18</td>
<td>(-2.3%, 12.3%)</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>32.4%</td>
<td>34.1%</td>
<td>1.6%</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bivariate Data</strong></td>
<td>Random.</td>
<td>58.1%</td>
<td>60.7%</td>
<td>2.6%</td>
<td>0.28</td>
<td>0.12</td>
<td>(-13.3%, 1.6%)</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>56.4%</td>
<td>64.8%</td>
<td>8.4%</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dordt’s before and after story

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Cohort</th>
<th>Pre</th>
<th>Post</th>
<th>Diff.</th>
<th>Paired t-test p-value</th>
<th>Cohort p-value</th>
<th>95% CI for cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob.</td>
<td>Random.</td>
<td>31.9%</td>
<td>56.5%</td>
<td>24.5%</td>
<td>&lt;0.001</td>
<td>0.52</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>32.4%</td>
<td>35.2%</td>
<td>2.7%</td>
<td></td>
<td></td>
<td>(10.8%, 32.7%)</td>
</tr>
<tr>
<td>Samp Var.</td>
<td>Random.</td>
<td>36.7%</td>
<td>39.4%</td>
<td>2.7%</td>
<td>0.22</td>
<td>0.11</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>38.7%</td>
<td>43.5%</td>
<td>4.8%</td>
<td></td>
<td></td>
<td>(-9.4%, 5.2%)</td>
</tr>
<tr>
<td>CIs</td>
<td>Random.</td>
<td>37.9%</td>
<td>51.8%</td>
<td>13.9%</td>
<td>&lt;0.001</td>
<td>0.12</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>Tradition.</td>
<td>42.9%</td>
<td>47.8%</td>
<td>4.9%</td>
<td></td>
<td></td>
<td>(1.1%, 16.7%)</td>
</tr>
<tr>
<td>Tests of</td>
<td>Random.</td>
<td>46.1%</td>
<td>70.0%</td>
<td>23.9%</td>
<td>0.000</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sig.</td>
<td>Tradition.</td>
<td>50.0%</td>
<td>60.6%</td>
<td>10.6%</td>
<td></td>
<td></td>
<td>(6.6%, 19.9%)</td>
</tr>
</tbody>
</table>
Transferability

- Fall 2013 and spring 2014
- 22 different instructor-semesters
- 17 different instructors
- 12 different institutions
- N=725; pre-post on 30 question ISI assessment (adapted from CAOS)

- Many different instructional styles (traditional classroom, active learning pedagogy, computer lab, flipped classroom)
- Many different institutions (high school, community college, large university, mid-sized university, small liberal arts college)
Transferability- Overall

- Similar findings to author’s institutions; Significantly better overall post-course performance
## Transferability – by subscale

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Pretest</th>
<th>Posttest</th>
<th>Diff.</th>
<th>Paired t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>48.7%</td>
<td>57.8%</td>
<td>9.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Data Collection and Design</strong></td>
<td>64.7%</td>
<td>67.2%</td>
<td>2.4%</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Descript. Statistics</strong></td>
<td>36.8%</td>
<td>44.5%</td>
<td>7.7%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Graphical Representations</strong></td>
<td>50.9%</td>
<td>59.0%</td>
<td>8.1%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Probability</strong></td>
<td>35.8%</td>
<td>47.2%</td>
<td>11.4%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Sampling Variability</strong></td>
<td>20.9%</td>
<td>24.8%</td>
<td>4.0%</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>CIs</strong></td>
<td>52.7%</td>
<td>64.2%</td>
<td>11.5%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Tests of Sig.</strong></td>
<td>58.7%</td>
<td>70.5%</td>
<td>11.8%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Low performers - overall

- Not leaving weak students behind; results similar to traditional curriculum
Discussion

- What we know
  - Anecdotal evidence growing; more and more people jumping on the bandwagon; sustained discussion, development of materials over the last decade
  - The ISI version of the curriculum (early, middle and current versions) have demonstrated
    - Improved learning gains in logic and scope of inference compared to traditional curriculum at same institutions
    - These results appear to translate reasonably well to other institutions---even those without direct comparison data
    - Improved retention of these same key areas
    - ‘Do no harm’ in descriptive statistics and other areas
    - Attitudes; conceptual/attitudes (Talk this afternoon; 1F1; Swanson)
Discussion

- What we don’t know
  - Pedagogy? Content? Spiraling?
    - Conflated!
  - What you should ‘take’ and what you can ‘leave’; student learning trajectories
  - Key instructor/institutional requirements for success
  - How the approach can be improved even further for greater success
Our plans...

○ Assessment initiative
  ○ Do you want to participate? Nathan.tintle@dordt.edu
  ○ Pre- and post- concepts and attitudes; common exam questions
  ○ ‘Non-users’ are especially needed!!
  ○ Goal: What works, what doesn’t, comparisons by institution, instructor, style, etc. Individualized instructor reports to learn about your own students outcomes

○ Dissemination of materials (prelim edition; other talks); continued refinement of materials; training on implementing randomization/simulation (workshop Saturday; JSM; more coming)

○ Continued conversation
  ○ Online community fall 2014
Other talks (among others)

- Swanson and VanderStoep
  - Attitudes; this afternoon 145PM; 1F1
- Chance and McGaughey.
  - More conceptual on specific areas
  - 6B1 (Thursday)
- Roy et al.
  - Overview of introduction of p-value in week 1.
  - 4A2 (Tuesday)
but more is needed

- Randomized experiments with targeted interventions to assess
  - particular student learning outcomes
  - effective pedagogical strategies and
  - to develop a clearer understanding of student development learning trajectories
Concluding analogy

○ Goal:
  ○ Give students a 360 degree view of statistical reasoning; a comprehensive understanding of description and inference; what statistics can and can’t tell

○ Are we there yet?
Option #1: Made it!

- We’ve blazed a trail to the top of the mountain; Randomization/simulation gives students the 360 degree view we want
Option #2: False summit

We thought we were almost to the top, but we’re not. We’re on the right route and climbing the right mountain, but not there yet. More work to do.
Option #3: Wrong mountain

- The only way to get higher is to go down and climb a different mountain
Option #4: Wrong continent?
Even if we have made it...

- We’re only halfway (we still have to get down!)
- Once we’re down we’ve got to figure out build a 4-lane highway to the top so we can bring the rest of the statistics education community with us
Acknowledgments

- Acknowledgments: ISI Team, other curriculum developers
- Funding: NSF (DUE-1140629 and DUE-1323210), Wiley, other funding agencies (HHMI; Teagle Foundation, etc.)
- Slides available at
  - http://math.hope.edu/isi (main textbook website)
References