Can the effectiveness of teaching methods be measured with final exam scores?

Georg Rieger
Physics & Astronomy, UBC
What is measured?

• ‘Performance’ score on final exam:

\[
\text{performance} = \frac{\text{exam percentage} \times \text{Bloom's level}}{\text{average Bloom's level (2.87)}}
\]

• Phys 100 (2006 – 2013)
• N = (640 – 840) students
• Style of final exam has not changed since 2006.
Fig. 1: Average final exam percentage and average final exam performance. Error bars reflect the standard deviation of the 2010 – 2014 data.
Table 1. Format of the final exams in Physics 100 and average scores. The number of multiple-choice N (MC) questions is shown in column 2; the number of parts in problem questions N (PQ) is shown in column 3. Columns 4 and 5 show the percentage weight of multiple-choice (MC %) and problem questions (PQ %) contributed to the final exam scores, respectively. The average exam score is in column 6 and the average Bloom’s level of each final exam is shown in column 7. The corresponding exam performance score = (Bloom average x Exam average)/(Average Bloom’s level) is shown in column 8.

<table>
<thead>
<tr>
<th>Year</th>
<th>N (MC)</th>
<th>N (PQ)</th>
<th>MC %</th>
<th>PQ %</th>
<th>Exam Average</th>
<th>Bloom’s Level</th>
<th>Performance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>10</td>
<td>15</td>
<td>40</td>
<td>60</td>
<td>59.7</td>
<td>2.75</td>
<td>57.2</td>
</tr>
<tr>
<td>2007</td>
<td>10</td>
<td>11</td>
<td>38</td>
<td>62</td>
<td>59.5</td>
<td>2.82</td>
<td>58.5</td>
</tr>
<tr>
<td>2008</td>
<td>9</td>
<td>10</td>
<td>47</td>
<td>53</td>
<td>54.4</td>
<td>3.11</td>
<td>58.9</td>
</tr>
<tr>
<td>2009</td>
<td>10</td>
<td>11</td>
<td>28</td>
<td>72</td>
<td>59.7</td>
<td>2.90</td>
<td>60.3</td>
</tr>
<tr>
<td>2010</td>
<td>16</td>
<td>9</td>
<td>50</td>
<td>50</td>
<td>64.8</td>
<td>2.80</td>
<td>63.2</td>
</tr>
<tr>
<td>2011</td>
<td>16</td>
<td>16</td>
<td>46</td>
<td>54</td>
<td>61.3</td>
<td>2.88</td>
<td>61.5</td>
</tr>
<tr>
<td>2012</td>
<td>15</td>
<td>16</td>
<td>38</td>
<td>62</td>
<td>62.4</td>
<td>2.81</td>
<td>61.1</td>
</tr>
<tr>
<td>2013</td>
<td>22</td>
<td>14</td>
<td>48</td>
<td>52</td>
<td>61.9</td>
<td>2.89</td>
<td>62.3</td>
</tr>
<tr>
<td>2014</td>
<td>22</td>
<td>14</td>
<td>48</td>
<td>52</td>
<td>56.3</td>
<td>3.04</td>
<td>59.6</td>
</tr>
</tbody>
</table>
Bloom’s Levels

- Evaluated by single rater (me)

- Two sources:
  - Bloom’s level chart with action words (from Carl’s learning goal presentation)
  - Blooming tool (Casagrand and Semsar, U of Colorado, unpublished)
Table 2. Column 2 shows the re-normalized performance = performance/(average Bloom’s level)*100. Columns 3 and 4 show CLASS results for pre-/posts shift in the general problem solving category and the overall shift, both for the favorable category. Column 5 shows the overall CLASS score (fav.) at the end of a term. The last column shows the new pedagogies introduced into the course. All new pedagogies are still in use. For example open-book exams are used since 2006. (Clickers and peer-instructions were introduced in 2002.) The CLASS data in columns 3 – 5 is corrected for the average grade dependence. {The result of the correction is shown in brackets.}

<table>
<thead>
<tr>
<th>Year</th>
<th>Normalized Performance</th>
<th>CLASS-PS_Shift (fav.) {adjusted}</th>
<th>CLASS-All_Shift (fav.) {adjusted}</th>
<th>CLASS-ALL_Post (fav.) {adjusted}</th>
<th>New Pedagogy</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>57.1</td>
<td>-5.5 ± 2.9 {-8.1}*</td>
<td>-2.7 ± 1.7 {-4.2}*</td>
<td>45.7 ± 2.0 {42.4}*</td>
<td>Open book midterm and final exams</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*Small sample (N=91); CLASS grade average very different from Course grade average (-7.8)</td>
</tr>
<tr>
<td>2007</td>
<td>58.5</td>
<td>0.5 ± 1.1 {-0.3}</td>
<td>-2.5 ± 0.7 {-2.7}</td>
<td>51.0 ± 0.9 {50.0}</td>
<td>Context-rich tutorials and group work; Learning Goals</td>
</tr>
<tr>
<td>2008</td>
<td>58.9</td>
<td>0.8 ± 1.2 {0.5}</td>
<td>-2.0 ± 0.8 {-2.2}</td>
<td>47.7 ± 0.9 {47.2}</td>
<td>Custom textbook</td>
</tr>
<tr>
<td>2009</td>
<td>60.3</td>
<td>-2.7 ± 1.2 {-3.0}</td>
<td>-5.4 ± 0.7 {-5.6}</td>
<td>47.4 ± 0.9 {47.0}</td>
<td>Pre-class reading assignments</td>
</tr>
<tr>
<td>2010</td>
<td>63.1</td>
<td>4.1 ± 1.4 {3.4}</td>
<td>-0.9 ± 0.9 {-1.4}</td>
<td>51.0 ± 1.1 {49.4}</td>
<td>Worksheets in lecture</td>
</tr>
<tr>
<td>2011</td>
<td>61.7</td>
<td>4.2 ± 1.1 {3.7}</td>
<td>0.5 ± 0.7 {0.1}</td>
<td>52.5 ± 0.9 {50.4}</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td>61.0</td>
<td>3.7 ± 1.4 {2.5}</td>
<td>-0.3 ± 0.8 {-0.9}</td>
<td>54.5 ± 1.0 {53.1}</td>
<td>Two-stage midterm exams</td>
</tr>
<tr>
<td>2013</td>
<td>62.4</td>
<td>No data</td>
<td>No data</td>
<td>No data</td>
<td></td>
</tr>
</tbody>
</table>
Analysis 2:
Another way to compare the data is to simply compare the averages and standard deviations for the (2006 – 2009) and (2010 – 2013) periods, which correspond to the years before and after introducing worksheets into the lecture portion. Table 3 shows the results.

**Table 3.** Average exam scores and performance scores aggregated for two time periods.

<table>
<thead>
<tr>
<th>Period</th>
<th>Exam Score</th>
<th>STD DEV</th>
<th>Performance</th>
<th>STD DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006 – 2009</td>
<td>58.3%</td>
<td>2.6%</td>
<td>58.7%</td>
<td>1.3%</td>
</tr>
<tr>
<td>2010 – 2013</td>
<td>62.6%</td>
<td>1.5%</td>
<td>62.1%</td>
<td>0.9%</td>
</tr>
</tbody>
</table>
Bloom's Taxonomy of the Cognitive Domain
(≈ content+skills+habits of mind)

1. Factual Knowledge: remember and recall factual information
   Define, List, State, Label, Name, Describe

2. Comprehension: demonstrate understanding of ideas, concepts
   Describe, Explain, Summarize, Interpret, Illustrate

3. Application: apply comprehension to unfamiliar situations
   Apply, Demonstrate, Use, Compute, Solve, Predict, Construct, Modify

4. Analysis: break down concepts into parts
   Compare, Contrast, Categorize, Distinguish, Identify, Infer

5. Synthesis: transform, combine ideas to create something new
   Develop, Create, Propose, Formulate, Design, Invent

6. Evaluation: think critically about and defend a position
   Judge, Appraise, Recommend, Justify, Defend, Criticize, Evaluate

Higher level: Require deeper conceptual understanding
Table 2. Bloom’s Dichotomous Key (BDK) (Casagrand and Semsar, U of Colorado)

- Categorize the question based on what students are being asked to do, not on how challenging the question may be. (For example, a 'comprehend' question for a difficult concept could be a more challenging problem than an 'analyze' question on an easier concept.)
- Evaluate questions with reference to what material we know students were exposed.

Q1. Could students memorize the answer to this specific question?
- Yes – Go to Q2.
- No – Go to Q4.

Q2. To answer the question, are students repeating nearly exactly what they have heard or seen in class materials (including lecture, textbook, lab, homework, clicker, etc.)?
- Yes → SEE RECALL
- No – Go to Q3.

Q3. Are students demonstrating a conceptual understanding by putting the answer in their own words, matching examples to concepts, representing a concept in a new form (words to graph, etc.), etc.?
- Yes → SEE COMPREHENSION
- No – Go BACK to Q1. If you are sure the answer to Q1 is yes, the question should fit into RECALL or COMPREHENSION.

Q4. Is there potentially more than one valid solution* (even if a “better” one exists, or if there is a limit to what solutions can be chosen)?
- Yes – Go to Q5.
- No – Go to Q8.

Q5. Are students making a judgment and/or justifying their answer?
- Yes → SEE EVALUATE
- No – Go to Q6.

Q6. Are students synthesizing information into a bigger picture (coherent whole) or creating something they haven’t seen before (a novel hypothesis, novel model, etc.)?
- Yes → SEE SYNTHESIZE/CREATE
- No – Go to Q7.

Q7. Are students being asked to compare/contrast information?
- Yes → SEE ANALYZE
- No - GO BACK through each category or refer category descriptions to see which fits the best.

Q8. To answer the question, do students have to interpret data (graph, table, figure, story problem, etc.)?
- Yes – Go to Q9.
- No – Go to Q14.

Q9. Are students determining whether the data are consistent with a given scenario or whether conclusions are consistent with the data?
- Yes → SEE EVALUATE

Q10. Are students building up a model or novel hypothesis from the data?
- Yes → SEE SYNTHESIZE/CREATE

Q11. Are students coming to a conclusion about what the data mean (they may or may not be required to explain the conclusion), and/or having to decide what data are important to solve the problem (i.e., picking out relevant from irrelevant information)?
- Yes → SEE ANALYZE

Q12. Are students using the data to calculate the value of a variable?
- Yes → SEE APPLY

Q13. Are students re-describing the data to demonstrate they understand what the data represent?
- Yes → SEE COMPREHEND

Q14. Are students putting information from several areas together to create a new pattern/structure/model/etc.?
- Yes → SEE SYNTHESIZE/CREATE

Q15. Are students predicting the outcome or trend of a fairly simple change to a scenario?
- Yes → SEE APPLY

Q16. Are students demonstrating that they understand a concept by putting it into a different form (new example, analogy, comparison, etc.) than they have seen in class?
- Yes → SEE COMPREHEND

*Note: The term “valid solution” refers to any answer that is logically consistent with the question and the data provided.