Robust Assessment Instrument for Student Problem Solving

Jennifer L. Docktor
Ken Heller, Leon Hsu

University of Minnesota
Physics Education Research & Development Group
http://groups.physics.umn.edu/physed
Problem Solving Measure

- Problem solving is both an important mechanism and outcome of learning.
  - This is certainly true in physics
  - Unfortunately, there is no standard way to easily measure problem solving so that student progress can be assessed.

- Our goal is to develop an easy-to-use, robust instrument to assess students' written solutions to physics problems, and obtain evidence for reliability, validity, and utility of scores.

- The instrument should be general
  - not specific to instructor practices or techniques
  - applicable to a range of problem topics and types
Reliability, Validity, & Utility

- **Reliability** – score agreement
- **Validity** evidence from multiple sources
  - Content (relevant & representative)
  - Response processes
  - Internal & external structure
  - Generalizability
  - Consequences of testing
- **Utility** - usefulness of scores


## Instrument at a Glance (Rubric)

### Separate score for each category – indicates strengths & weaknesses

### Minimize the number of categories & scores

<table>
<thead>
<tr>
<th>CATEGORY</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
<th>NA (Prob)</th>
<th>NA (Slvr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physics Approach</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specific Application</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math Procedures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logical Progression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SCORE**
Research Basis for Rubric

Representative research literature

- **Categories come from:**
  - **Problem solving processes**
  - **Expert-novice characteristics**
  - **Previous work at Minnesota**

- **Instrument construction:**
  - **Validity, Reliability, Utility**
  - **Rubrics**
### Rubric Scores (in general)

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Complete &amp; appropriate</td>
<td>Minor omission or errors</td>
<td>Parts missing and/or contain errors</td>
<td>Most missing and/or contain errors</td>
<td>All inappropriate</td>
<td>No evidence of category</td>
</tr>
</tbody>
</table>

**NOT APPLICABLE (NA):**

<table>
<thead>
<tr>
<th>NA - Problem</th>
<th>NA - Solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not necessary for this problem (i.e. visualization or physics principles given)</td>
<td>Not necessary for this solver (i.e. able to solve without explicit statement)</td>
</tr>
</tbody>
</table>
Overview of Study

1. Drafting the rubric
2. Preliminary tests with two raters (final exams and instructor solutions)
3. Training exercise with graduate students
4. Analysis of tests from an introductory mechanics course
5. Student problem-solving interviews (in progress)
Initial Training Exercise

Results - preliminary rubric use

8 physics graduate students with TA experience

- Score agreement improved significantly with minimal training
  - Weighted kappa $0.27 \pm 0.03$ Fair $\rightarrow 0.42 \pm 0.03$ Moderate
  - Math & Logical Progression most affected
- Raters influenced by traditional grading experience
  - Unwilling to score math and logic if physics incorrect
- Multi-part problems more difficult to score
- Revisions to rubric and training based on this
  - Consistent language across rubric categories
  - More examples of NA scores & expand zero score
  - Distinguish physics approach & application

\[
\kappa_w = \frac{\sum w_{ij}f_{oi} - \sum w_{ij}f_{Eij}}{(w_{\max})N - \sum w_{ij}f_{Eij}}
\]

Cohen’s Weighted Kappa
Analysis of Tests

- Calculus-based introductory physics course for Science & Engineering students (mechanics)
- Test problems graded in the usual way by teaching assistants, then scored with rubric by researcher

EXAMPLE DATA

**Instructor Solutions**

Professor Solutions to Textbook Problems
Calculus-Based Mechanics Homework

**Student Solutions**

FREQUENCY OF RUBRIC SCORES
CALCULUS-BASED PHYSICS FOR SCIENCE & ENGINEERING, FALL 2008 EXAM 3 PROB 1

- MOST ERRORS IN SPECIFIC APPLICATION OF PHYSICS
- MATH GENERALLY OK
- Useful Description
- Physics Approach
- Specific Application
- Math Procedures
- Logical Progression

SCORE FREQUENCY (N=188)

RUBRIC SCORE

0 1 2 3 4 5 NA(P) NA(S)
Findings from Test Analysis

- The rubric indicates areas of student difficulty for a given problem
  - Focus instruction to coach physics, math, clear and logical reasoning processes, etc.

- The rubric responds to different problem features
  - Can make score interpretation difficult
    - usefulness of visualization
    - prompts & cues
    - numeric vs. symbolic question
  - Modify problems to elicit / practice processes
Summary

- A rubric is being developed from descriptions of problem solving process, expert-novice research studies, and past studies at UMN
  - Focus on written solutions to physics problems
- Training revised to improve score agreement
- Rubric provides useful information that can be used for research & instruction
- Rubric works for standard range of physics topics in an introductory mechanics course
  - There are some problem characteristics that make score interpretation difficult (prompts & cues)
- Interviews will provide information about response processes
For more information:

docktor@physics.umn.edu
http://groups.physics.umn.edu/physed
Additional Slides
References - Categories


References – Instrument Construction


Weighted Kappa

\[
\kappa_w = \frac{\sum w_{ij} f_{oij} - \sum w_{ij} f_{Eij}}{(w_{\text{max}}) N - \sum w_{ij} f_{Eij}} = \frac{\sum f_{o(w)} - \sum f_{E(w)}}{(w_{\text{max}}) N - \sum f_{E(w)}}
\]

\[
\sigma_{\kappa_w} = \sqrt{\frac{N \sum w_{ij}^2 f_{oij} - (\sum w_{ij} f_{oij})^2}{N (w_{\text{max}} N - \sum w_{ij} f_{Eij})^2}}
\]

\[
z = \frac{\kappa_{w1} - \kappa_{w2}}{\sqrt{\sigma_{\kappa_{w1}}^2 - \sigma_{\kappa_{w2}}^2}}
\]

95% confidence limit: 1.960

99% confidence limit: 2.576

99.9% confidence limit: 3.291

fo: observed frequencies of exact agreement (diagonal of pivot table)

fe: expected frequencies of exact agreement by chance

N: number of subjects rated
Kappa

\[ \kappa = \frac{\sum f_o - \sum f_E}{N - \sum f_E} \]

\[ \sigma_{\kappa} = \sqrt{\frac{N \sum f_o - (\sum f_o)^2}{N(N - \sum f_E)^2}} \]

\[ z = \frac{\kappa_1 - \kappa_2}{\sqrt{\sigma_{\kappa_1}^2 - \sigma_{\kappa_2}^2}} \]

- \( f_o \): observed frequencies of exact agreement (diagonal of pivot table)
- \( f_E \): expected frequencies of exact agreement by chance
- \( N \): number of subjects rated

95% confidence limit: 1.960
99% confidence limit: 2.576
99.9% confidence limit: 3.291