Moneyball
Are your students getting on base?

Planning and Research Office
Craig Hayward, Ph.D.
Natalia Cordoba-Velasquez
Cabrillo College
January 31st, 2012
First, a primer on baseball:

- [http://www.youtube.com/watch?v=cMha-DjYMqQ](http://www.youtube.com/watch?v=cMha-DjYMqQ)
What is sabermetrics?

- The search for objective knowledge about baseball
- The value of getting on base

*Baseball statistics, unlike statistics in any other area, have acquired the power of language.*

- Bill James, 1985 Statistical Abstract
The four inefficiencies

1. Not basing decisions on data
2. Using the wrong data to make decisions
3. Using good data but in the wrong way
4. Not collecting the right data
The **Moneyball** odyssey

- Where did Bill James start?
  - He started where the data was the best
  - “Looking at places where the stats don’t tell the whole truth – or even lie about the situation.”
### See a pattern?

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Credit Grade</strong></td>
<td><strong>Credit Grade Count (%)</strong></td>
<td><strong>Credit Grade Count (%)</strong></td>
<td><strong>Credit Grade Count (%)</strong></td>
<td><strong>Credit Grade Count (%)</strong></td>
<td><strong>Credit Grade Count (%)</strong></td>
</tr>
<tr>
<td>Cabrillo Total</td>
<td>38,954</td>
<td>100.00%</td>
<td>38,289</td>
<td>100.00%</td>
<td>38,339</td>
</tr>
<tr>
<td>Grade A</td>
<td>10,602</td>
<td>27.22%</td>
<td>10,603</td>
<td>27.69%</td>
<td>11,651</td>
</tr>
<tr>
<td>Grade B</td>
<td>6,083</td>
<td>15.62%</td>
<td>6,121</td>
<td>15.99%</td>
<td>6,013</td>
</tr>
<tr>
<td>Grade C</td>
<td>3,568</td>
<td>9.16%</td>
<td>3,327</td>
<td>8.69%</td>
<td>3,025</td>
</tr>
<tr>
<td>Grade D</td>
<td>815</td>
<td>2.09%</td>
<td>746</td>
<td>1.95%</td>
<td>754</td>
</tr>
<tr>
<td>Grade F</td>
<td>695</td>
<td>1.78%</td>
<td>830</td>
<td>2.17%</td>
<td>974</td>
</tr>
<tr>
<td>Pass</td>
<td>6,591</td>
<td>16.92%</td>
<td>5,405</td>
<td>14.12%</td>
<td>5,856</td>
</tr>
<tr>
<td>No Pass</td>
<td>2,513</td>
<td>6.45%</td>
<td>2,604</td>
<td>6.80%</td>
<td>3,048</td>
</tr>
<tr>
<td>Incomplete No Credit</td>
<td>575</td>
<td>1.48%</td>
<td>730</td>
<td>1.91%</td>
<td>413</td>
</tr>
<tr>
<td>Report Delayed</td>
<td>0.00%</td>
<td>0.00%</td>
<td>54</td>
<td>0.14%</td>
<td>125</td>
</tr>
<tr>
<td>Dropped</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Withdrew</td>
<td>6,326</td>
<td>16.24%</td>
<td>6,999</td>
<td>18.28%</td>
<td>5,535</td>
</tr>
<tr>
<td>Military Withdrawal</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Unknown</td>
<td>1,186</td>
<td>3.04%</td>
<td>870</td>
<td>2.27%</td>
<td>945</td>
</tr>
</tbody>
</table>

### Credit Grade Count (%)

- **Spring 1993**: 38,954
- **Spring 1995**: 38,289
- **Spring 1997**: 38,870
- **Spring 1999**: 38,339
- **Spring 2001**: 38,159

### Grade Distribution

- **Grade A**: 10,602 (27.22%)
- **Grade B**: 6,083 (15.62%)
- **Grade C**: 3,568 (9.16%)
- **Grade D**: 815 (2.09%)
- **Grade F**: 695 (1.78%)
- **Pass**: 6,591 (16.92%)
- **No Pass**: 2,513 (6.45%)
- **Incomplete No Credit**: 575 (1.48%)
- **Report Delayed**: 54 (0.14%)
- **Dropped**: 0.00%
- **Withdrew**: 6,326 (16.24%)
- **Military Withdrawal**: 0.00%
- **Unknown**: 1,186 (3.04%)
Proportion of “A” Grades relative to all other grade notations

y = 0.0052x + 0.2721
R² = 0.8349

Grade inflation?
A bit less inflated

\[ y = 0.0041x + 0.3774 \]

\[ R^2 = 0.4208 \]
Grade inflation in proper context

\[ y = 2E-05x + 0.5076 \]

\[ R^2 = 0.00002 \]
Cabrillo College’s Campus Climate Study

- Biennial survey
- Major revision in 2008
  - Dropped demographic questions
  - Collected data sufficient for a “fuzzy match”
  - Added engagement, behavioral & tech questions
- 2,055 cases from Fall 2008 & Fall 2010
# Sample Description

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>College Population Fact Book 2011</th>
<th>Campus Climate Sample (n=2,055)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Female</td>
<td>53%</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>Latino</td>
<td>30%</td>
</tr>
<tr>
<td></td>
<td>White</td>
<td>57%</td>
</tr>
<tr>
<td>Age</td>
<td>18-20</td>
<td>29%</td>
</tr>
<tr>
<td></td>
<td>21-25</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>26-30</td>
<td>12%</td>
</tr>
<tr>
<td>Workload</td>
<td>Full Time</td>
<td>28%</td>
</tr>
</tbody>
</table>
Student engagement is “…the interaction or fusion of behavior, emotion, and cognition in the process of learning.”

Student Engagement

- Participated in class: 2.92
- Rapid instructor feedback: 2.77
- Asked instructor re: assignments: 2.75
- Had meaningful conversations with students of different ethnicity: 2.54
- Worked with other students: 2.52
- Sought advice re: career plans: 2.05
- Used a chat or email for class: 2.00
Full Time students are engaged

(statistically significant differences for all items)
Technology usage 2008

Frequency of Usage 2008

Facebook
- Never: 62%
- Seldom: 10%
- Sometimes: 12%
- Often: 16%

My Space
- Never: 39%
- Seldom: 19%
- Sometimes: 19%
- Often: 23%

C. Wireless
- Never: 63%
- Seldom: 12%
- Sometimes: 14%
- Often: 11%
Building a model of student achievement

- **Multivariate Linear Regression**
  - Does the inclusion of a factor change the model?
  - Standardized Beta coefficients range from -1 to 1

- **Dependent Variable:** GPA

- **16 Independent/predictor variables tested**

- **Hypothesis** – student engagement has a direct effect on student achievement
Is the influence of Student Engagement on achievement mediated by other factors or does it have a direct effect?

- **Student Engagement**
- **Demographics**
  - Age
  - Gender
  - Ethnicity
  - SES
  - Other factors
- **Student Achievement (GPA)**
The basic relationship

Student Engagement \[ r = .11 \] GPA
Unit Load and Working

- Unit Load
- Hours Working
- Interaction of load*work

R = .15

GPA
## Unit Load and Working - details

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>(Constant)</td>
<td>3.040</td>
<td>.146</td>
<td>20.867</td>
</tr>
<tr>
<td></td>
<td>Hours worked</td>
<td>.083</td>
<td>.031</td>
<td>.201</td>
</tr>
<tr>
<td></td>
<td>Term Units</td>
<td>-.004</td>
<td>.012</td>
<td>-.018</td>
</tr>
<tr>
<td></td>
<td>Work*units interaction</td>
<td>-.006</td>
<td>.003</td>
<td>-.197</td>
</tr>
</tbody>
</table>

a. Dependent Variable: GPA
Demographics

- Gender
- Age
- Ethnicity

R = .33

GPA
Demographic model - details

<table>
<thead>
<tr>
<th>Model</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.507</td>
<td>.063</td>
<td></td>
<td>40.070</td>
<td>.000</td>
</tr>
<tr>
<td>(Constant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.022</td>
<td>.002</td>
<td>.249</td>
<td>10.273</td>
<td>.000</td>
</tr>
<tr>
<td>Gender</td>
<td>.165</td>
<td>.041</td>
<td>.096</td>
<td>4.001</td>
<td>.000</td>
</tr>
<tr>
<td>Latino</td>
<td>-.265</td>
<td>.043</td>
<td>-.149</td>
<td>-6.192</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: GPA

N.B. Age has a bivariate association with GPA of virtually zero! $r = .022$
Uber model

- Age
- Ethnicity
- Gender
- Teacher support
- Unit load
- Home Tech Live with parents
- Engagement

$R = .39$

GPA
<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>2.375</td>
<td>.181</td>
<td>13.100</td>
<td>.000</td>
</tr>
<tr>
<td>Age</td>
<td>.017</td>
<td>.003</td>
<td>.191</td>
<td>5.924</td>
</tr>
<tr>
<td>Gender</td>
<td>.151</td>
<td>.045</td>
<td>.089</td>
<td>3.326</td>
</tr>
<tr>
<td>Latino</td>
<td>-.184</td>
<td>.049</td>
<td>-.105</td>
<td>-3.775</td>
</tr>
<tr>
<td>Perception of Instructors</td>
<td>.086</td>
<td>.030</td>
<td>.080</td>
<td>2.860</td>
</tr>
<tr>
<td>Technology in Home</td>
<td>.084</td>
<td>.031</td>
<td>.074</td>
<td>-2.716</td>
</tr>
<tr>
<td>Term Units</td>
<td>-.014</td>
<td>.006</td>
<td>-.069</td>
<td>-2.399</td>
</tr>
<tr>
<td>Living with parents</td>
<td>-.168</td>
<td>.052</td>
<td>-.100</td>
<td>-3.218</td>
</tr>
<tr>
<td>engagement</td>
<td>.100</td>
<td>.037</td>
<td>.076</td>
<td>2.698</td>
</tr>
</tbody>
</table>

a. Dependent Variable: GPA
What opportunities are we missing?

- What data do you think might be important to predict achievement that we are not currently collecting/using?
Next Steps

- Continue to reflect on how to use predictive information
  - In what contexts can this information be used to enhance student success?

- Integrating psychological measures
  - the College Self-Assessment Survey (CSSAS)
  - Research question: Do psychological measures enhance our ability to predict student performance?

- Consider benefits of integrating Climate survey with Instructional Planning survey
CSSAS CONSTRUCTS:

Academic Self-Efficacy  
Hope  
Academic Identity  
Goals  
Personal Responsibility  

Self-Regulation  
Relationship to Self  
Relationship to Others  

Achievement (GPA)  

Interventions
• Learning communities  
• Grant activities  
• Curricular innovation  
• Matched comparison groups  

Communication  
Leadership & Teamwork
The four inefficiencies

1. Not basing decisions on data
   ◦ “Death by anecdote”
2. Using the wrong data to make decisions
   ◦ Granularity; Simpson’s paradox
3. Using good data but in the wrong way
   ◦ Grade inflation
4. Not collecting the right data
   ◦ Missed classes, missed opportunities
The answers I arrive at – and thus the methods that I choose – are almost never wholly satisfactory, never wholly disappointing. The most consistent problems that I have arise from the limitations on my information sources.

Bill James as quoted in Moneyball, page 82
Mauriello and Armbruster’s goal was to value the events that occurred on a baseball field more accurately than they had ever been valued before. In 1994, they stopped analyzing derivatives and formed a company to analyze baseball players, called AVM Systems. Ken Mauriello had seen a connection between the new complex financial markets and baseball: “the inefficiency caused by sloppy data.” As Bill James had shown, baseball data conflated luck and skill, and simply ignored a lot of what happened in a game. – Moneyball, page 131