

OUTCOMES

*Assessing
Student Performance
And
Measuring Clerkship Success*

DISCLOSURE

Nothing to disclose

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

TWO KINDS OF OUTCOMES

- *How are students doing?*
- *How good is the clerkship?*

WHY MEASURE OUTCOMES?

*The **GOOD** Reasons*

- How are students doing?
 - Does **this student** know enough neuro to be a good doctor?
 - How can this student improve?
- How good is the clerkship?
 - Do **all students** know enough neuro to be good doctors?
 - How can the clerkship be improved?
- NOTE: For these outcomes, *no need to rank students*

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

The Fundamental Problem: All Outcome Measures are Surrogates

- How are students doing?
 - Does **this student** know enough neuro to be a good doctor?
 - How can this student improve?

- How good is the clerkship?
 - Do **all students** know enough neuro to be good doctors?
 - How can the clerkship be improved?

*The Fundamental Problem:
All Outcome Measures are Surrogates.
They may be ...*

- Irrelevant
- Inaccurate or Imprecise
- Incomplete
- Multifactorial

WHY MEASURE OUTCOMES?

*The **OTHER** Reasons*

- How are students doing?
 - How does this student compare to others?
 - Cum Laude, AOA, etc.
 - Residency rankings
- How good is the clerkship?
 - Distribution of med school/dep't. funds
 - Faculty assignments, promotions and awards

THE BOTTOM LINE

- Outcome measures are a necessary evil
 - *(just remember to keep them in perspective)*

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Written*
 - *Oral*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Written*
 - *MCQ*
 - *Short Answer*
 - *Essay*
 - *Oral*
 - *Clinical evaluations*

- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

WRITTEN EXAMS

MCQ vs Short Answer/Essay

- Advantages of MCQ:
 - Easy to grade
 - Standardized
 - Students are used to them and need to keep in practice
- Advantages of Short Answer/Essay
 - Provide a better sense of student thought processes
 - More realistic
 - Easier to develop the tests

The Winner: MCQ Exams NBME vs Locally Developed

- *Advantages of NBME*
 - *Professionally prepared and edited*
 - *Statistical power*
 - *External (national) norms*
 - *Less work*
- *Advantages of Locally-developed exam*
 - *Control over content, format, and length*
 - *Flexibility*
 - *Cheaper (except start-up costs)*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Written*
 - *Oral*
 - *Structured (OSCE: Objective Structured Clinical Examination)*
 - *Unstructured*
 - *Clinical evaluations*

- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

Oral Exams: Pros & Cons

■ *Pros*

- *Allow real-time interaction and probing of student understanding*
- *Questions can be modified based on student performance*
- *Potential for assessing clinical skills with patients (real or standardized)*

■ *Cons*

- *Time-consuming*
- *Impossible to be completely realistic*
- *Hard to standardize*

OSCEs: Pros & Cons

■ *Pros*

- *Standardized (by definition)*
- *Potential for assessing clinical skills with patients (standardized)*
- *Allow real-time interaction and probing of student understanding*

■ *Cons*

- *Even more time-consuming and expensive than unstructured oral exams*
- *Still not totally realistic*
- *Standardization limits flexibility*
- *Standardization highlights “buzzwords”*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
 - *Summative*
 - *Fragmentary*

- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

Summative Clinical Evaluations

Typical Format

- *Component scores*
 - *History*
 - *Physical Exam*
 - *Oral presentations*
 - *Write-ups*
 - *Clinical reasoning*
 - *Dependability*
 - *Etc.*
- *Overall Score*
- *Rating scale 1 – N (N = 5-10)*
- *Descriptors (1 means ...)*
- *Guidelines (what percentage of students receive 4-6 ...)*

DEPARTMENT OF NEUROLOGY M3 CLERKSHIP EVALUATION FORM

UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN

Student's Name: _____
 Evaluator's Name: _____
 Hospital & Service: _____
 Period (Month): _____

At most 10% of students will fall in these categories: At least 50% of students will fall in these categories: 30% of students will fall in these categories: 10% of students will fall in these categories:

- | | | | |
|--|---|--|---|
| 1=Severe deficiencies | 3=One or two minor deficiencies, otherwise meets basic standard | 5=Consistently exceeds basic standards | 7=The top 10% of his/her class. |
| 2=Less than adequate performance, several significant deficiencies | 4=Meets all basic standards, may exceed them at times | 6=Consistently exceeds basic standards, many notable strengths | 8=One of the best 10 students I've ever worked with |

Please evaluate the student's performance for each component of clinical competence. Circle the rating which best describes the student's skills and abilities. Use as your standard the level of skill expected from the clearly satisfactory student at this stage of training. Identify (by circling relevant phrases and/or providing separate comments on the reverse side) strengths and weaknesses you have observed. For any component that **needs attention** or you are unable to judge due to **insufficient contact** with the student, please check the appropriate category. Be as specific as possible, including reports of critical incidents. Global adjectives or remarks, such as "good student," do not provide as meaningful feedback to the student as specific comments.

1. MEDICAL HISTORY AND INTERVIEW

Incomplete, inaccurate information. Important information often missing. Unreliable, superficial.	1	2	3	4	5	6	7	8	Precise, thorough, reliable information. Logical and comprehensive.
<input type="checkbox"/> Needs attention									
<input type="checkbox"/> Insufficient contact to judge									

2. PHYSICAL EXAMINATION

Incomplete, inaccurate, cursory. Unable to interpret findings.	1	2	3	4	5	6	7	8	Complete, accurate, directed at patient's problems. Elicits even subtle findings.
<input type="checkbox"/> Needs attention									
<input type="checkbox"/> Insufficient contact to judge									

3. MEDICAL RECORDS

Inaccurate, incomplete, disorganized, verbose.	1	2	3	4	5	6	7	8	Accurate, complete, logical and concise.
<input type="checkbox"/> Needs attention									
<input type="checkbox"/> Insufficient contact to judge									

4. ORAL PRESENTATIONS

Rambling, inaccurate, incomplete, disorganized.	1	2	3	4	5	6	7	8	Concise, accurate, complete and orderly.
<input type="checkbox"/> Needs attention									
<input type="checkbox"/> Insufficient contact to judge									

5. MEDICAL KNOWLEDGE

Fragmented, limited; unable to recall basic science and clinical information.	1	2	3	4	5	6	7	8	Extensive, well integrated base of pertinent basic science and clinical knowledge.
<input type="checkbox"/> Needs attention									
<input type="checkbox"/> Insufficient contact to judge									

PLEASE COMPLETE REVERSE SIDE ALSO

**University of Michigan Medical School
Clinical Performance Assessment
Third Year Required Clerkships 2005-2006**

10% of students:

1=Very poor performance, many deficiencies
2=Poor performance
3=Marginal performance

80% of students:

4=Less than average performance
5=Average performance; appropriate for level
6=Above average performance

10% of students:

7=Performance in top 10% of class
8=One of the best 10 students I've worked with
9=Best student I've ever had

1. Medical History and Interview

Incomplete, inaccurate information.
Missing important information.
Unreliable, superficial.

Precise, thorough, reliable information.
Logical and comprehensive.
Well-organized.

1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Needs attention

2. Physical Examination

Incomplete, inaccurate, cursory.
Unable to interpret findings.

Complete, accurate, directed at patient's problems.
Elicits subtle findings, able to interpret findings.

1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Needs attention

3. Medical Records

Incomplete, inaccurate,
disorganized, verbose.

Accurate, complete,
logical, and concise.

1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Needs attention

4. Oral Presentations

Summative Clinical Evaluations

Typical Problems

- *The “halo” effect (e.g., a ‘6’ in every category)*
- *Grade inflation*
- *Student concerns*
 - *Faculty vs. senior residents vs. junior residents*
 - *Early in year vs. late in year*
 - *Subjectivity*

Summative Clinical Evaluations

Typical Problems

- *The “halo” effect (e.g., a ‘6’ in every category)*
- *Grade inflation*
- *Student concerns*
 - *Faculty vs. senior residents vs. junior residents*
 - *Early in year vs. late in year*
 - *Subjectivity*

Results

19 faculty members rated students both pre and post

	No Feedback	Feedback
Mean scores	5.35	5.54
Tightness of Scoring	1.09	1.05
Score deviance	0.13	-0.02
ABS deviance	0.21	0.20

Summative Clinical Evaluations

Typical Problems

- *The “halo” effect (e.g., a ‘6’ in every category)*
- *Grade inflation*
- *Student concerns*
 - *Faculty vs. senior residents vs. junior residents*
 - *Early in year vs. late in year*
 - *Subjectivity*

Evaluations: Seniority Effects?

■ METHODS

- All residents who
 - started in 1996-7, 7-8, 8-9, 9-0, or 0-1
 - completed evals in all 3 years of training
- All faculty who had worked with the same students

Evaluations: Seniority Effects?

RESULTS

– 23 residents, 41 faculty, 914 students

MEAN CLINICAL SCORES ASSIGNED

<u>H.O. 2</u>	<u>H.O. 3</u>	<u>H.O. 4</u>	<u>Faculty</u>
5.85	5.78	5.92	5.68

No statistically significant differences

Summative Clinical Evaluations

Typical Problems

- *The “halo” effect (e.g., a ‘6’ in every category)*
- *Grade inflation*
- *Student concerns*
 - *Faculty vs. senior residents vs. junior residents*
 - *Early in year vs. late in year*
 - *Subjectivity*

Effects of Timing on Neurology Clerkship Clinical Scores

	1 st Qtr	2 nd Qtr	3 rd Qtr	4 th Qtr
# studs	356	360	331	346
MAX	6.8	7.0	6.8	7.0
MEAN	5.5*	5.8*	5.7	5.7
MIN	4.1	4.3	4.3	4.3

* p = .035

Summative Clinical Evaluations

Typical Problems

- *The “halo” effect (e.g., a ‘6’ in every category)*
- *Grade inflation*
- *Student concerns*
 - *Faculty vs. senior residents vs. junior residents*
 - *Early in year vs. late in year*
 - *Subjectivity*

Results

19 faculty members rated students both pre and post

	No Feedback	Feedback
Mean scores	5.35	5.54
Tightness of Scoring	1.09	1.05
Score deviance	0.13	-0.02
ABS deviance	0.21	0.20

13. Overall Clinical Performance

1	2	3	4	5	6	7	8	9
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

SCALE CHANGES

14. History Taking and Physical Examinations

Overall, this student is currently functioning at this level:

Not yet ready for M3 year	Early M3 year	Mid M3 year	Late M3 year	M4/Subintern	Intern
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

15. Clinical Judgment and Decision Making

Overall, this student is currently functioning at this level:

Not yet ready for M3 year	Early M3 year	Mid M3 year	Late M3 year	M4/Subintern	Intern
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Please list student's strengths (required):

Please list areas where student needs improvement (required):

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
 - *Summative*
 - *Fragmentary*

- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

Fragmentary Clinical Evaluations – Examples

- *Review of a physical examination*
- *Forms for outpatient use:*
 - *Localization*
 - *Likely diagnosis*
 - *Diagnostic plan*
 - *Management plan*
- *Review of oral presentation*
- *Review of write-up*
- *Weekly evaluation*

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

Using Student Performance to Assess Clerkship Success

Duh

Using Student Performance to Assess Clerkship Success

- *NBME scores*
 - *Provide standardization and a national norm, but do they measure what's important?*
- *Patient outcomes*
 - *These **are** what's important, but don't usually reflect student (or clerkship) performance*
- *Patient comments*
 - *Useful, but not consistent*
- *Performance in residency and beyond*
 - *This is what's **truly** important, but*
 - *Hard to measure*
 - *Too many intervening variables*

Gelb's Law of Outcome Measures

- The more straightforward the educational outcome measure, the more tenuous the relationship to the ultimate desired outcome

COMMON OUTCOME MEASURES

- *How are students doing?*
 - *Exams*
 - *Clinical evaluations*
- *How good is the clerkship?*
 - *Student performance*
 - *Student ratings of clerkship/teachers*
 - *External (internal) review*

Students' Ratings of Teaching: Is the customer always right?

- Are the evaluations consistent?
 - no: for any random pair of students
 - yes: if at least 15 students (and at least 2/3 of the group) complete evaluations
- Are the evaluations stable over time?
 - yes

Students' Ratings of Teaching: Is the customer always right?

- Is there a halo effect?
 - no: factor analysis shows evaluations are multidimensional (but the number and nature of factors vary from one study to another)
- Global score and subscores may have different uses

Students' Ratings of Teaching: Is the customer always right?

- Do students' ratings of teachers correlate with students' test scores?
 - Yes, but:
 - not always
 - the magnitude of correlation is not dramatic

Limitations of Correlation Studies

- Must control for possibility that student test scores influence their ratings of teachers
 - Tests should be external
 - Ratings should be from students not in the study
- Teacher's skill is not the only factor in learning
 - Therefore, must:
 - either randomize student assignment
 - or control for all relevant variables
- Test scores do not measure all learning (especially clinical)

Searches for Confounding Variables

- Dr. Fox paradigm
 - expressiveness: large effect on ratings, small effect on achievement
 - content: the opposite
- But:
 - results vary
 - expressiveness IS important for effective communication

Searches for Confounding Variables - 2

- Students' grade expectations DO correlate with their ratings of instructors
 - can be interpreted in various ways
- Students' ratings DO NOT correlate with:
 - instructor's age, sex, race, teaching experience, or research productivity
 - student's age, sex, GPA, or personality

Searches for Confounding Variables - 3

Learning objectives – the Gelb experience

Students' Evaluations of Teaching: Summary

- Consistent and representative (as long as at least 15 responses, at least 2/3 of students)
- Correlate (but not strongly) with other measures of teaching effectiveness
- Correlate with instructor expressiveness and expected grades
- Conclusion: useful, as long as limitations recognized

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

THE BOTTOM LINE

- "Everybody does it:"
 - *Some MCQ exam (NBME or home-grown)*
 - *A summative evaluation form*
 - *Student evaluations of clerkship and teachers*
- Optional extras (if you're ambitious):
 - *Short answer or essay exams*
 - *Oral exams (unstructured or OSCE)*
 - *Fragmentary evaluations*
 - *Internal or external reviews*

AGENDA

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*

ASSIGNING GRADES

$$(2/3 \times \text{Clinical}) + (1/3 \times \text{Exam})$$

ASSIGNING GRADES

$$(2/3 \times \text{Clinical}) + (1/3 \times \text{Exam})$$

Since exam scores are almost all > 70, and none has ever been below 50, expand range:

$$E^* = (E - 50) \times 2$$

ASSIGNING GRADES

$$(2/3 \times \text{Clinical}) + (1/3 \times ((E - 50) \times 2))$$

Since exam scores are 0-100, and clinical scores are 1-9, normalize exam score:

$$\begin{aligned} E^{**} &= (E^*/100) \times 9 \\ &= E^* \times .09 \end{aligned}$$

$$\text{So: } E^{**} = (E - 50) \times 2 \times .09 = (E - 50) \times .18$$

ASSIGNING GRADES

$$(2/3 \times \text{Clinical}) + (1/3 \times ((E - 50) \times 18))$$

- Each evaluator's clinical score is calculated by taking the mean of all component scores, then averaging with the overall score:

$$C^* = ((C1 + C2 + \dots + C12)/12 + C)/2$$

- **The clinical scores from all evaluators (faculty and neurology residents) are averaged, weighting each evaluator's score by the number of weeks on service with the student**

ASSIGNING GRADES

Each member of the grading committee independently assigns a letter grade to each student and submits these grades to the clerkship director

The clerkship director then assigns the final letter grade: H, HP, P, or F (majority rule, rounding up in case of ties)

Students who fail the exam but do well on clinical evaluations can remediate by taking a make-up examination

ONE LAST POINT:

Measure baselines

Figure 1: Exam Scores Before and After Period 19

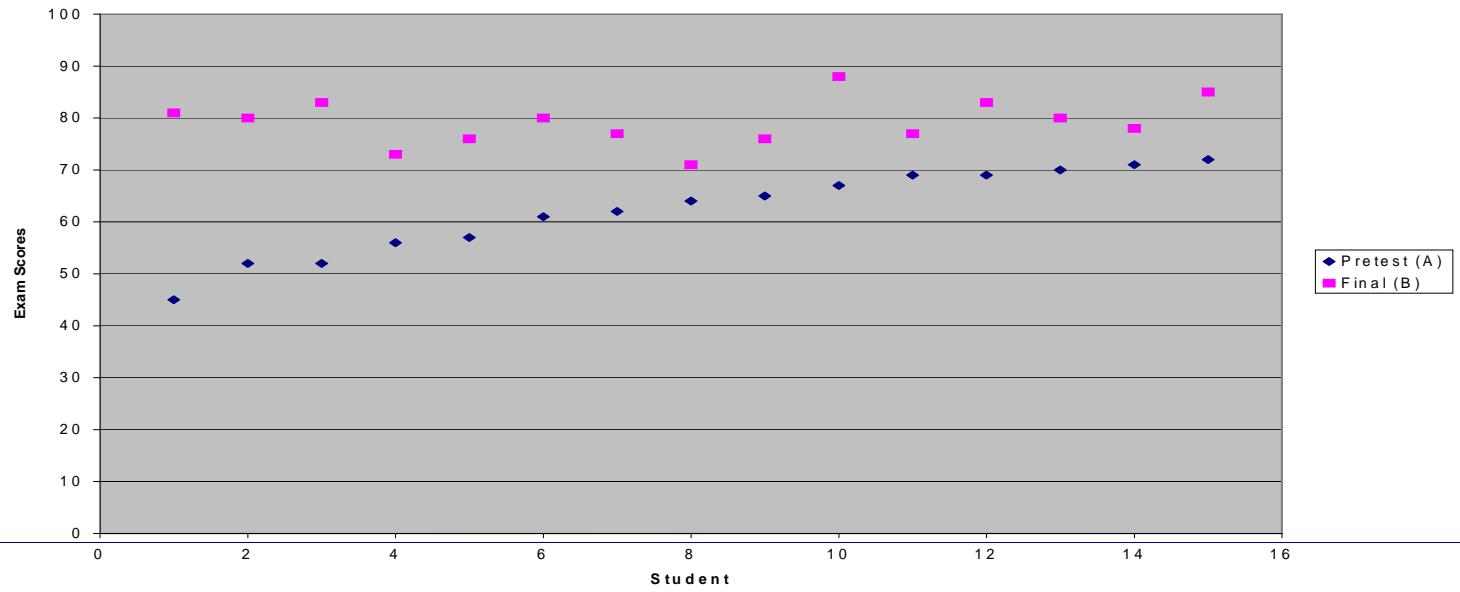
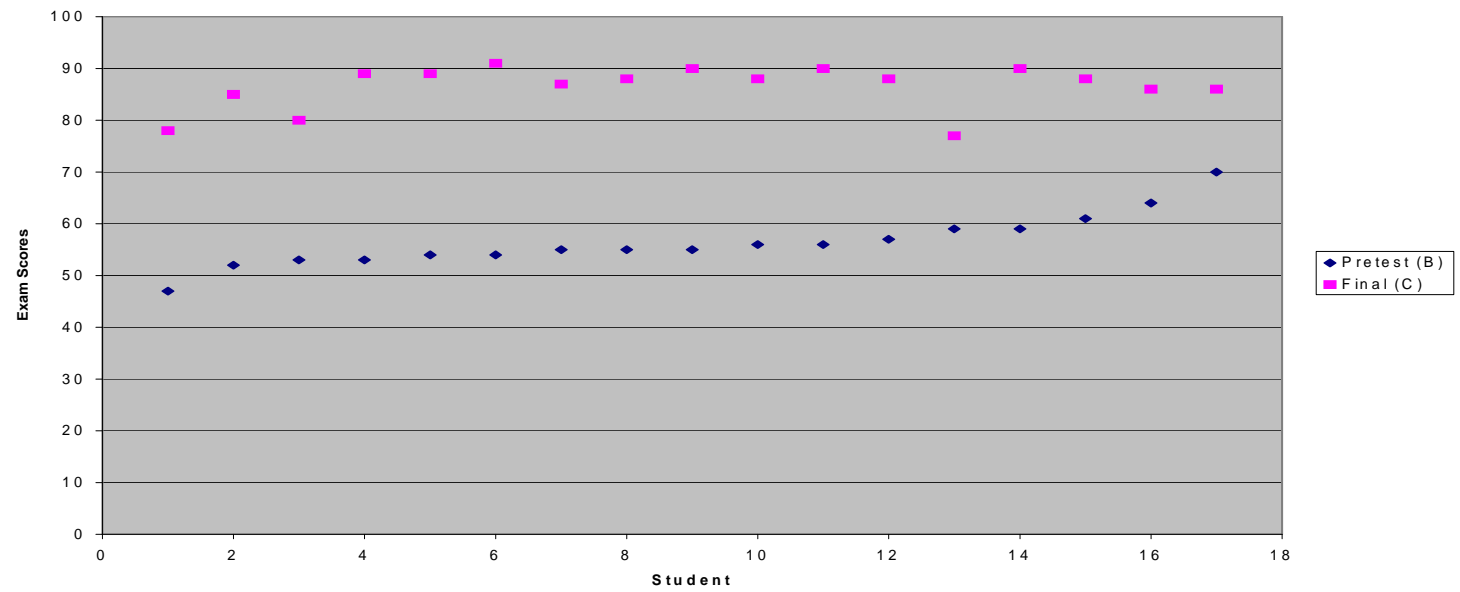


Figure 2: Exam Scores Before and After Period 20



RECAP

- *A Little Philosophy*
 - *Why measure outcomes?*
 - *The downside*
- *Some Brass Tacks*
 - *Common outcome measures*
- *Brassier and Tackier*
 - *Assigning grades*