Do Grades Impede Learning?

Gender in STEM Education: Tools for Change

Brown Bag Lunch Talk—Wednesday March 17th, 2010

Morgan Benton

Agenda

- Not here to sell you on something, but to provoke thought
- Introduce 3 thought-provoking papers
  - Vickers (2000)—Justice and Truth in Grades and Their Averages
  - Moss (2003)—Reconceptualizing Validity for Classroom Assessment
  - Kohn (2002)—The Dangerous Myth of Grade Inflation
- Relate my experiences with "going gradeless"
- Discussion

My Background

- 5 years teaching English in rural Japanese middle school
- 9 years full-time teaching university-level programming
- Graduate research intern at ETS, Summer 2005
- Dissertation Title (2008): The Development and Evaluation of Software to Foster Professional Development in Educational Assessment
- 3 years as a CFI TAP consultant at JMU

Why examine grades?

- Because we expend a great deal of time and energy
  - Thinking about them
  - Devising ways to calculate them
  - Negotiating with students about them
  - Recording, safeguarding, documenting them
- Because almost no one ever does…

Key Questions

- What, if anything, do grades measure?
- If grades are measurements, what arguments can be made to support their reliability, accuracy, precision, and validity?
- Regardless of measurement issues, do grades, on balance, impede or promote learning?

Justice and Truth—Vickers (2000)

- Grading practices:
  - Are not uniform
  - Don't distinguish difficulty (encourage gaming the system)
  - Are insensitive to the number of courses taken
  - Don't distinguish skills, e.g. A+C vs. C+A vs. B+B
  - Sometimes involve "corrective" weighting schemes
- GPA = Abstraction of an Abstraction
- Goal of grades → Preserve and Transmit Information
- Goal of Paper: Examine structural properties of grade averaging

Assumptions—Vickers (2000)

- Grades are not relative to teachers or evaluators
- Grades objectively measure the quality of student work,
  i.e. A work is in fact better than B work which is better than C work
- Properties:
  - Transitivity: If $X > Y$ and $Y > Z$ then $X > Z$; If $X=Y$ and $Y=Z$ then $X=Z$
  - Asymmetry: If $X > Y$ then $Y \sim> X$
  - Reflexivity: $X = X$
  - Symmetry: If $X = Y$ then $Y = X$
  - Connexity: $X > Y$ or $Y > X$ or $X = Y$

Scales Vary—Vickers (2000)

Typical 4-point Scale

A	B	C	D	F
4	3	2	1	0

CGU Eight-Point Scale (Ramified)

A+	A	A-	B+	B	B-	C	C-	U
8 (4.0)	7 (4.0)	6 (3.7)	5 (3.3)	4 (3.0)	3 (2.7)	1 (2.0)	(1.7)	0

Multiple Scales in use at Same School

Level	A	B	C	D	F
I	4.00	3.00	2.00	1.00	0
II	5.00	4.00	3.00	1.50	0
III	6.00	5.00	4.00	2.00	0
IV	7.00	6.00	5.00	2.50	0
V	8.00	7.00	6.00	3.00	0

Ramifications—Vickers (2000)

- Goal of GPA is clear → rank ordering of students
- Benign case:
  - Student A: A A A A
  - Student B: A A A C
- Contradictory Case:
  - Student C: A A F
  - Student D: C C C

Scale	Student			
	A	**B**	**C**	**D**
CGU8	**7.0**	5.8	**4.67**	1.00
CGU(r)	**4.0**	3.6	**2.67**	2.00
I	**4.0**	3.6	**2.67**	2.00
II	**5.0**	4.6	**3.33**	3.00
III	**6.0**	5.6	4.00	4.00
IV	**7.0**	6.6	4.67	**5.00**
V	**8.0**	7.6	5.33	**6.00**

Assessment Validity—Moss (2003)

- Asks the question: What does it mean for classroom assessment of learning to be "valid?"
- Contrasts 2 definitions of learning: psychometric and socio-cultural
- Explores what it means to do assessment when one adopts the socio-cultural definition of learning

Psychometric Definition—Moss (2003)

- Learning is characterized by what we can infer from observed changes in individuals' performance on assessments over time
- This view dominates our educational culture
    - Virginia Standards of Learning Tests (SOLs)
    - No Child Left Behind (NCLB)
    - Prevalence of tests/quizzes in classroom assessment

Socio-Cultural Definition—Moss (2003)

"From a sociocultural perspective, learning is perceived through changing relationships among the learner, the other human participants, and the tools (material and symbolic) available in a given context. Thus learning involves not only acquiring new knowledge and skill, but taking on a new identity and social position within a particular discourse or community of practice. As Wenger puts it, learning 'changes who we are by changing our ability to participate, to belong and to experience our life and the world as meaningful'." (p.14)

I couldn't have said it better…—Moss (2003)

"Informal consideration of interactional evidence with these sorts of questions in mind helped me make the decision to abandon grades, whenever possible. I had always found the giving of grades to require a substantial commitment of time to develop a meaningful rubric and assign scores fairly-time that took me away from tasks that seemed to have a higher pedagogical value. I began to attend more explicitly to how they shaped my interactions with students about their work, both before and after the assignment of the grade. Conversations too frequently focused on what I wanted, on what I considered necessary for an A, or on why a higher grade than the one I had assigned was fair. When I gave students opportunities to revise their work to improve the grade or I postponed the giving of a grade until revised versions were turned in, I found the revision typically accomplished just what I had asked for and nothing more. Ungraded rubrics functioned in much the same fashion. As Shepard (2003) notes: "competitive grading practices seem to be so pervasive in US. classrooms that the purpose of rubrics has been corrupted from criteria that make the features of excellent work accessible into a point system used for defending and quarreling over grades" (p. 176). I don't want the capital in my classroom to be grades or even my approval; it will not sustain students (as professionals) outside the classroom. I want it to be doing something that is meaningful and useful within the context of classroom and the relevant research communities." (p 19)

Dangerous Myth of Grade Inflation—Kohn (2002)

- *Grade inflation got started … in the late '60s and early '70s.… The grades that faculty members now give … deserve to be a scandal.*

  -- Professor Harvey Mansfield, Harvard University, 2001

- *Grades A and B are sometimes given too readily -- Grade A for work of no very high merit, and Grade B for work not far above mediocrity. ... One of the chief obstacles to raising the standards of the degree is the readiness with which insincere students gain passable grades by sham work.*

  -- Report of the Committee on Raising the Standard, Harvard University, **1894**

The Argument—Kohn (2002)

- Cries of “grade inflation” are rarely if ever accompanied by either data or reasoned argument
  - Hard to substantiate that grades are rising
  - Even if grades have risen, what makes them undeserved?
  - Implicit argument about (reduced) accuracy of grading
  - Learning almost *never* enters the discussion
  - Economics, inputs, and outputs is the dominant metaphor

Unpacking the Myth—Kohn (2002)

- Premises underlying complaints about grade inflation
  - Professor's job is to sort students
  - Grades provide useful information to post-college constituencies
  - Students should be forced to compete for artificially scarce rewards
  - A normal distribution indicates "rigor" ("…*rather it is a symbol of failure—failure to teach well, failure to test well, and failure to have any influence at all on the intellectual lives of students.*")
  - Harder is better; confounding difficulty with quality
  - Scarcity of A's makes students work harder; grades motivate

What the Data Says—Kohn (2002)

- Grades may not be rising
- Changes in grading practices may explain differences
- Grades undermine motivation

Going Gradeless

- Action research paradigm
- 2 Semesters
  - Spring 2009—3 sections, total of 54 students
    - Required, intro-level programming course for non-programming majors
  - Fall 2009—1 section, 25 students
    - Elective, 2nd course in programming, database, and web application development
- The Story
- The Results

Key Questions

- What, if anything, do grades measure?
- If grades are measurements, what arguments can be made to support their reliability, accuracy, precision, and validity?
- Regardless of measurement issues, do grades, on balance, impede or promote learning?

Beliefs and Values

- Mastery Learning: Every student can and should succeed
- Social Constructivism:
  - Each student must define success; though I offer guidance
  - Comparing students to one another is inappropriate
- My primary role is educator—not credential-giver or HR rep
- Lifelong commitment trumps amount of content covered
- Grades hurt

Theoretical Foundation

- The pedagogy is grounded in Self Determination Theory, which posits that students have three basic needs:
  - Relatedness
  - Competence
  - Autonomy
- These are the foundation for fostering intrinsic motivation for learning course content

Relatedness

- Students want to feel a sense of relatedness to each other, to the content, and to the instructor
- This is fostered with:
  - Teams from day one
  - Hacking sessions
  - Relinquishing my role as judge
  - Incorporating reflection into labs

Competence

- Students need to experience challenge and success often; there's no better motivator than "getting it"
- This is achieved by:
  - Making it okay to fail; creating safe spaces for risk taking
  - Devoting an entire class each week to in-class, peer evaluation
  - Allowing students to dictate the pace of the course
  - Providing a variety of resources for building skill, e.g. videos, in-class tutorials, a knowledgeable TA, and yes, the text

Autonomy

- Students need to feel that they have control over their lives, that the things they care about can be a part of their classes
- This is accomplished by:
  - Making **everything** optional
  - Supporting them in challenging projects of their own choosing
  - Constantly reminding them that they are in control
  - Constantly asking them why they have made certain decisions

Accountability

- No grades ≠ No accountability
  - We call you from class when you don't show up
  - We may visit you if you miss repeatedly (3+ times in a row)
  - We thank slackers publicly
  - Your team and other teams count on you during every class

    … and most importantly …
  - We hold a mirror up to your face and ask you constantly to evaluate yourself

Structure

- I still drive the bus. I still command the bully pulpit. I'm still the one wearing the pants in this family.
  - Students have little (if any) experience making educational decisions; they need some guidance (but not too much)
  - A solid weekly rhythm provides a comfortable boundary
  - Short, clear labs set minimal expectations, but inspire students to push the boundaries of their comfort zones
  - There are still ≈12 labs, 3 exams, and 1-2 projects

Psychometric Learning

- Difference in observed performance on assessments
- Example:

Test Score 2:	94
Test Score 1:	- 80
Learning:	14

- The field of psychometrics is entirely devoted to ensuring that "14" is a meaningful number
- Psychometrics is the source of Classical Test Theory and Item Response Theory

Classical Test Theory

- Goal: Be scientific about the types of questions used to develop tests of human abilities
- Key Concept: **Item Discrimination**—the ability of any particular test item to discriminate between people of high and low ability on the given skill
- Classic calculation: $D_i = \dfrac{U_i - L_i}{U_i}$
- Interpretation:
  - High values (closer to 1) indicate good discrimination
  - Low values (closer to 0) indicate poor discrimination
  - Negative values indicate problem question

Problems

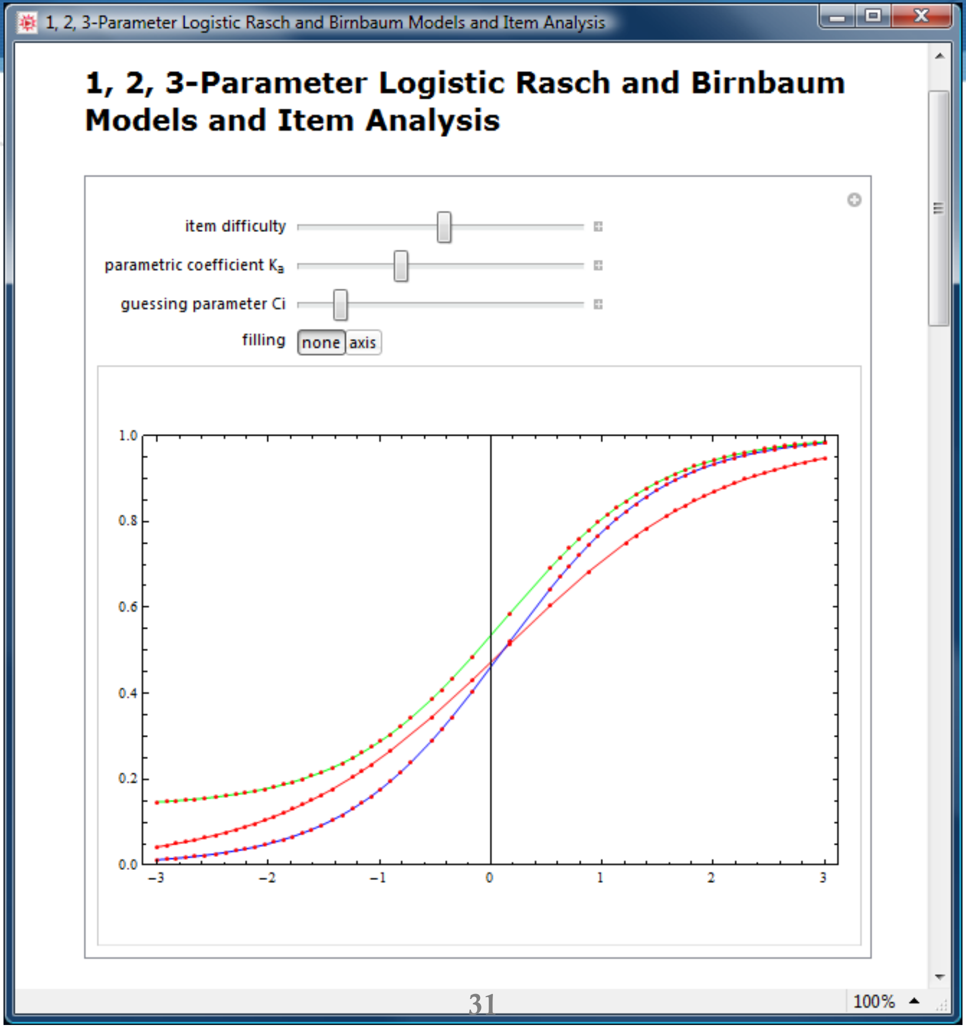
- The performance of items is dependent on:
  - The particular set of test takers
  - The particular set of questions chosen
- No way to select questions reliably
- Item Response Theory overcomes these shortcomings allowing us to build valid and reliable objective models of item performance as expressed in Item Characteristic Curves

The 3-Parameter Model

$$P(\theta) = c + \frac{(1-c)}{1+\exp[-1.7a(\theta-b)]}$$

- *ϑ*: ability of the test taker on skill being assessed
- *P(ϑ)*: probability of a correct answer given *ϑ*
- *a*: the discrimination parameter
- *b*: the difficulty parameter
- *c*: the pseudochance (“guessing”) parameter

Sample ICCs



1, 2, 3-Parameter Logistic Rasch and Birnbaum Models and Item Analysis

1, 2, 3-Parameter Logistic Rasch and Birnbaum Models and Item Analysis

item difficulty

parametric coefficient K_a

guessing parameter Ci

filling [✓ none] axis

Problems with IRT

- Developing valid and reliable items is ***extremely*** labor intensive because the parameters are unknown *a priori*
- All parameters in the models must be estimated using techniques like joint maximum likelihood estimation (joint MLE) or Bayesian procedures
- All estimation procedures require 500-1000 responses on any given item before parameters can be estimated for the 3-parameter model
- As such, use of IRT for validation of assessment items in classroom settings is impractical

Is IRT-Quality Reliability Necessary?

- If trying to assess students’ ability to read an interpret code, is there anything wrong with this problem?

Determine the output of the following code segment when the Start button is clicked:

```
 Private Sub Special(ByRef ASingle as Single, ByVal BSingle as Single)
  Dim CSingle As Single

  ASingle = 2 * ASingle
  BSingle = BSingle + 2
  CSingle = CSingle + 1
  OutTextBox.Text += ASingle.ToString + BSingle.ToString + CSingle.ToString + vbNewLine
End Sub

Private Sub StartButton_Click (...) Handles btnStart.Click
  Dim XSingle As Single, YSingle As Single

  XSingle = 2
  YSingle = 3
  Call Special(XSingle, YSingle)
  OutTextBox.Text += XSingle.ToString + YSingle.ToString + vbNewLine
  Call Special(XSingle, YSingle)
  OutTextBox.Text += XSingle.ToString + YSingle.ToString + vbNewLine
End Sub
```

Answer: Clearly Not

- ***For purposes of instruction*** it is clear that the assessments we devise are good enough to guide our pedagogical decisions and to provide students with enough information to make sound choices with respect to their own learning
- ***But…*** there are problems when we try to come up with scores that are then reported to actors outside of the classroom