May 2007

Toward a Common Currency: Using Item Response Theory to Adjust for Grading Inconsistency

Daniel W. Allen
Macalester College, danielallen725@gmail.com

Toward a Common Currency:
Using Item Response Theory to Adjust for Grading Inconsistency

Daniel Allen

Advisor: Danny Kaplan
Mathematics Department

Submitted: April 30, 2007

Abstract:

Grading inconsistency results from the application of different standards such that grades provide an inaccurate, and sometimes misleading, evaluation of student performance. Many attempts to correct for grading inconsistency have either not accurately corrected for inconsistency or have been too complicated to constitute politically feasible solutions to the problem. Within this context, Item Response Theory (IRT) is offered as a solution to grading inconsistency. IRT corrects for inconsistencies in the distributions yielded by the various grading practices of professors and departments as well as the failure of current practices to accurately account for differing student quality by course.

Table of Contents

I. Introduction

> "For students, the currency of academia is the grade. As the only tangible benefit that students receive for performing well in their courses, grades provide the primary mechanism available to faculty for maintaining academic standards. In a very real sense, professors pay grades to students in return for mastery of course material, and students barter these grades for jobs or entrance into professional or graduate school.
>
> "In the early 1960's, student grades, like the dollar, were taken off the gold standard. Over the course of the next thirty-five years, inflation led to significant decreases in the value of both college grades and the dollar….It [grade inflation] damaged the system by drawing attention away from a much more serious flaw in the academic monetary system: It masked wild fluctuations in the exchange rates between academic departments." (Johnson 2003; 196)

This project will deal with grading inconsistency. Grading inconsistency results from the application of different standards (whether by professor or department) such that grades provide an inaccurate, and sometimes misleading, evaluation of student performance. The problem of grading inconsistency is related to, but not the same as, the more commonly debated, hot-button issue of grade inflation. Grade inflation is defined as a general upward trend over the years in grades at colleges and universities nationwide that is frequently considered as reflective of lowering academic standards. Grade inflation is not synonymous with grading inconsistency because if professors uniformly raised (or "inflated") grades, then consistency in grading practices would be preserved. However, in practice this does not occur; while some faculty members proceed to give out higher student grades, other faculty members will adamantly maintain previous grading levels (or even become more stringent in reaction). And, whether the issue is grade inflation or grading inconsistency—the core concern is the same; namely, that grades become so unreliable as to be meaningless.

Thus, the debates are intricately related—it would be difficult to discuss grading inconsistency without simultaneously addressing grade inflation. However, the focus of this

paper will be on resolving grading inconsistency. If grades were only inflated, they could still constitute accurate indicators of relative student performance with the application of some fairly simple adjustments. However, when inconsistency is added to the equation, acquiring an accurate understanding of a student's academic performance becomes far more difficult.

Grading inconsistency is not entirely a bad thing. Grades are assigned to serve a number of functions; they may be used to assess the quality of a student's performance, provide students with motivation to complete coursework, and to encourage or discourage students from taking a particular course. Given this myriad of objectives, it is hardly surprising that professors often choose to strike a different balance and assign grades on differing bases. Indeed, a tremendous advantage of grades is the flexibility with which they provide professors in shaping the atmosphere of individual courses.

Nevertheless, if left unadjusted, such grading inconsistency has negative consequences. There are three reasons why this problem warrants attention: it affects students during their time at the college, it affects students when they leave school for the workforce or graduate school, and it affects the ability of teachers to instruct effectively. Accordingly, this paper will seek to develop a method whereby student's grades with different professors or in different departments can reasonably be compared.

This paper will proceed as follows: Section II reviews previous scholarship which has demonstrated the existence of grading inconsistency at some schools. Section III discusses the impact of grading inconsistency and why it is an issue worth remedying. Section IV examines proposals which have been forwarded to address the issue and explains why each of them either fails to accurately correct for grading inconsistency or is too complicated mathematically to constitute a politically feasible policy. Section V explains Item Response Theory and argues that it constitutes the best approach to address grading inconsistency.

Section VI provides a simulation using data from Macalester College to demonstrate the efficacy of one grade adjustment method. Section VII concludes and recommends that Macalester include an adjusted class rank based upon Item Response Theory.

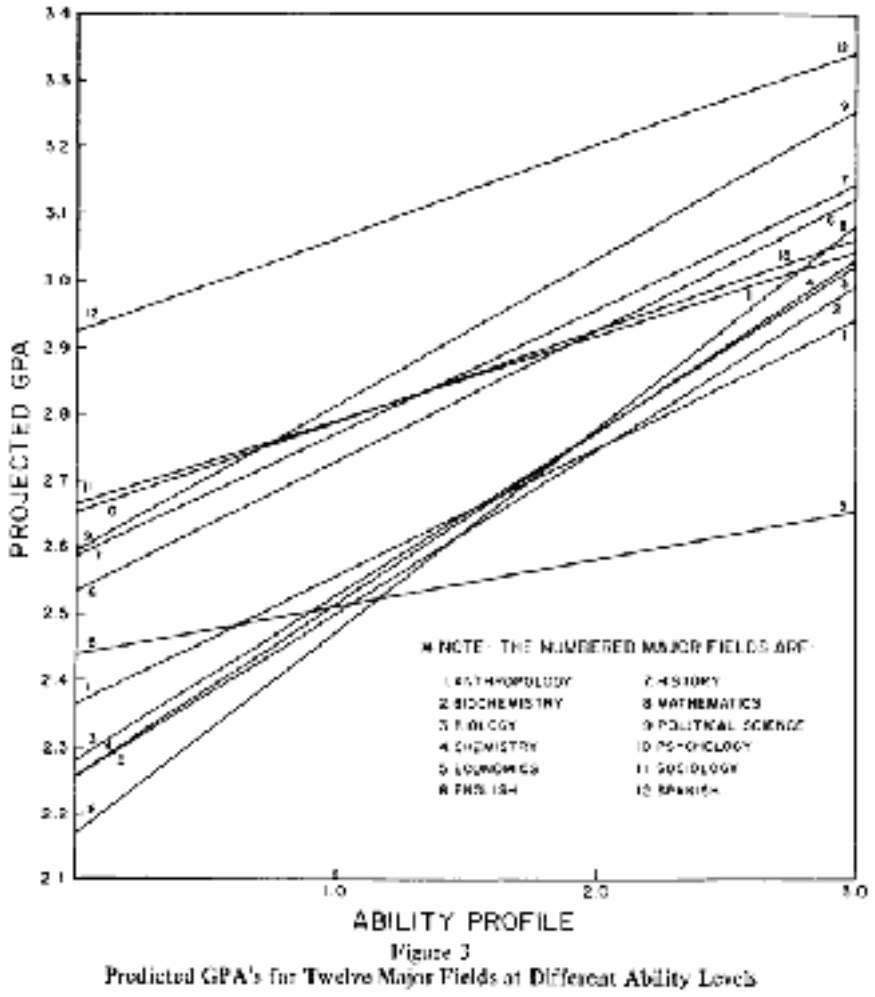
II. Studies Establishing the Existence of Grading Inconsistency

Within the literature on grading inconsistency, two basic methodologies are employed to establish its existence—those which are exogenous, relying on external measures of student ability, and those which are endogenous, relying exclusively upon grades themselves. The initial investigations of the topic relied upon exogenous measures such as standardized test scores and high school GPA.

Goldman, Schmidt, Hewitt and Fisher (1974) used three variables (SAT Math score, SAT Verbal score and high school GPA) to predict the grade point of students taking courses in different disciplines at the University of California, Riverside, during the 1972-1973 academic year. From the three predictor variables, Goldman *et. al.* derive a single variable, which they refer to as the student's "ability profile". Predicted GPA by department is then plotted against this ability profile (all fitting was done using linear regression). The resulting plot showed a series of lines separated at times by as much as 0.8 units of projected GPA (on a 4.0 scale; see Figure 1).

Spanish emerged as the most lenient grading of the departments. The History, Psychology and Sociology departments also graded more leniently than average. At the other end of the spectrum, the Economics, Anthropology, Biochemistry, Chemistry and Mathematics departments were the more severely grading departments.

Figure 1: Translation of ability profile to projected GPA by department for students at the University of California, Riverside, during the 1972-1973 academic year



Source: Goldman, Roy D., Schmidt, Donald E., Hewitt, Barbara Newlin, Fisher, Ronald. (1974), "Grading Practices in Different Major Fields." *American Educational Research Journal*, Vol. 11, No. 4 (Autumn), p. 343-357.

While it is possible to observe general trends such as the above, precise statements on the relative difficulty of grading in different departments are not possible due to the non-parallel nature of the lines translating "ability profile" to projected GPA. The Economics line, for example, is relatively flat, meaning that either the department does a poor job distinguishing between students or that prior measured abilities have little effect on performance in the discipline. On the other extreme, the line for Mathematics is quite steep, meaning that the department does distinguish well between students and that high ability

students (at least as indicated by high school GPA and SAT score) are likely to perform better.

Subsequently, studies have tended to focus more on endogenous measures of grading inconsistency. Exogenous measures are subject to severe limitations. High school GPA is of limited utility because high school students are competing with a much broader population; students at the same university likely represent a small sliver of that broad high school population and are therefore likely to have similar GPAs. Thus, although high school curriculum is more standardized than at the college level, it will fail to sort well between like-ability students. Moreover, there is substantial inconsistency in grading between high schools which makes comparison difficult. SAT, on the other hand, aims at too low a cognitive level to accurately reflect the ability of top students and has come under heavy scrutiny for potential racial and gender biases (Young 2003; Leonard and Jiang 1999). Ultimately, exogenous measures are limited because they are only as meaningful as the measures they rely upon. Thus, the shortcomings of high school GPA and the SAT as ability measures represent shortcomings in exogenous measures of ability.

The basis for the endogenous literature on grading inconsistency is work done by Goldman and Widawki (1976). In their paper, Goldman and Widawski were the first authors to employ what they term a "with-in subjects technique" to determine the extent of grading inconsistency. The technique used data from the spring quarter of 1973 at the University of California, Riverside, to construct pairwise comparisons of grades received by the same student within different departments. If a student had taken courses in two different departments, the student's average grade in each department was determined. The resulting difference was combined with those of other students, and the unweighted mean of the group was then taken as the net grading bias between the two departments. The authors

compensate for the tendency of individual strengths and weaknesses to effect the direction and magnitude of the resulting differential by aggregating data from a large number of students spread throughout the college.

Goldman and Widawski's analysis produced two important findings. The first finding was that grading inconsistency did, in fact, exist. The second finding was that the grade differentials between departments were transitive; thus, if Biology is 0.2 tougher than Chemistry which is 0.3 tougher than Physics, then Biology will be approximately 0.5 tougher than Physics. This property allowed the authors to construct a "Grading Index" centered at zero, where more difficult departments had negative coefficients and the more lenient departments had positive coefficients (see Table 1).

Table 1: Grading Index showing the degree of grading inconsistency at the University of California, Riverside during the Spring quarter of 1973

Combined Major Fields	**Grading Index**
Anthropology	-.09
Art	.22
Biology	-.53
Chemistry	-.36
Economics	.18
English	-.10
Ethnic Studies	.38
Foreign Language	.06
Geology	-.30
History	.05
Mathematics	-.07
Philosophy	.17
Physics	-.23
Political Science	-.03
Psychology	.03
Sociology	.29
Urban Studies	.38

Source: Golman, Roy D. and Widawski, Mel H. (1976), "A Within-Subjects Technique for Comparing College Grading Standards: Implications in the Validity of the Evaluation of College Achivement." *Educational and Psychological Measurement*, Vol. 36, p. 381-390.

This transitivity is not trivial. It is possible to construct a very simple scenario where it would not hold, as shown in Table 2. While this type of scenario unfolds in individual cases, the mean behavior of the student body still balances out such that department differentials are transitive.

Table 2: Basic scenario where grade transitivity does not hold

Student	Biology 1	Chemistry 1	Physics 1
Al	A		B
Beth	B	A	
Charlie		B	A

The methodology employed by Goldman and Widawski has its limitations. To demonstrate the largest shortcoming, let's consider a pairwise comparison between the Art and Mathematics departments. In doing the comparison, the method does not consider whether the students taking both courses tend to be Mathematics or Art majors. If, hypothetically, Art majors were far more likely to take Mathematics courses than Mathematics majors were to take Art courses, and if we assume students perform better within their major, the Mathematics courses are likely to appear to be more difficult than Art courses. This concern is particularly relevant given the different skill sets which the two departments draw upon. This shortcoming would later be remedied by Johnson (2003), who considered the direction of the comparison and found that grading discrepancies persisted. Despite its limitations, the study represented an advance because it relied exclusively upon grade data (instead of on external indicators of student ability such as standardized test scores and high school GPA).

A series of follow-up studies was done at Dartmouth College during the 1980's based upon the method used by Goldman and Widawski (1976). The first study, by A. Christopher Strenta and Rogers Elliot (1987), used data from the graduating class of 1983.

Their results confirmed those of Goldman and Widawski: substantial differentials between departments emerged and these differentials once again proved to be additive. A subsequent study by the same authors (Elliot and Strenta, 1988) confirmed these findings (see Table 3) and suggested remedying departmental biases through an additive adjustment. This remedy will be considered later, when grade adjustment methods are addressed.

Table 3: *Department adjustment constants for the Dartmouth class of 1986*.

(*Note*: Positive constants represent more leniently grading departments, while negative coefficients represent more strictly grading departments).

Department	Additive Constant
Anthropology	.00
Art/Visual Studies	.06
Asian Studies	.12
Biology	-.32
Chemistry/Biochemistry	-.35
Comparative Literature	.31
Drama	.37
Earth Science	.02
Economics	-.44
English	.05
Engineering	-.16
French & Italian	.08
Geography	-.16
German	-.07
Government	-.19
Greek & Roman Studies	.05
History	-.07
Math/Computer Science	-.37
Music	.29
Philosophy	-.07
Physics	-.25
Policy Studies	.09
Psychology	-.16
Religion	.03
Russian	.02
Sociology	.25
Spanish	.20

Source: Elliot and Strenta. Rogers and A. Christopher. 1988. "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly." *Journal of Educational Measurement*, Vol. 25, No. 4. (Winter). p. 333-347.

The non-parallel nature of the lines found in Goldman *et. al.* (1974) demonstrates an important limitation of the studies done by Strenta and Elliot (1987) and Elliot and Strenta (1988). Their comparison of disciplines assumes that an additive adjustment is sufficient to remedy the differences in grading practices across departments. However, the work of Goldman *et al.* (1974) demonstrates that both a multiplicative and an additive component would be necessary to translate different grades to a 'common currency' or single uniform scale reflecting relative student achievement.

In all of these studies a common thread emerges. Courses in economics, mathematics and the sciences are graded more stringently, courses in the humanities tend to be graded more leniently, and courses in the social sciences (except for economics) tend to lie somewhere in the middle. These general trends appear to hold at Macalester as well. Table 4 shows average grades by department at Macalester over the last five years.

The figures within this table are troubling. An average student in Chemistry, for example, will likely rank in the tenth percentile of the graduating class[1]. An average student in Economics will likely rank in the fifteenth percentile of the graduating class and an average student in Biology will rank in the twentieth percentile of the graduating class.

While individuals may have differing perceptions of the strength of students in various departments, I would venture to say that few, if any, would seriously make the claim that the mathematics, science and economics students at Macalester and other universities consistently prove to be the poorest students by such a large margin. And, even if there were no significant difference in grading practices by department, there would still exist differences in practice by professor. This inconsistency alone is sufficient to justify some form of adjustment.

[1] Data comes from the Macalester College Factbook produced by Institutional Research, which provides the distribution of graduating grade point averages.

Table 4: Grade Point Average by department at Macalester, Spring 2002-Spring 2006.

Department	Spring 2002	Spring 2003	Spring 2004	Spring 2005	Spring 2006
Neuroscience and Cognitive Studies	-	-	-	3.80	3.91
Education	3.72	3.71	3.72	3.60	3.72
American Studies	-	-	-	3.40	3.69
Russian Studies	3.48	3.27	3.43	3.42	3.64
English	3.50	3.58	3.52	3.46	3.63
History	3.40	3.48	3.39	3.40	3.61
Geography	3.42	3.57	3.52	3.48	3.60
Anthropology	3.55	3.57	3.45	3.54	3.59
Art	3.57	3.57	3.63	3.66	3.59
Music	3.36	3.43	3.58	3.58	3.58
Theatre and Dance	3.52	3.64	3.33	3.48	3.57
Asian Studies	3.49	3.58	3.74	3.86	3.55
Japanese Program	3.37	3.49	3.49	3.41	3.51
Religious Studies	3.51	3.65	3.56	3.61	3.51
International Studies	3.34	3.23	3.38	3.30	3.50
Psychology	3.39	3.44	3.52	3.47	3.50
Environmental Studies	3.69	3.30	3.57	3.39	3.48
Political Science	3.55	3.41	3.59	3.48	3.48
Sociology	3.55	3.57	3.50	3.48	3.48
Women's & Gender Studies	3.54	3.51	3.43	3.38	3.48
Hispanic & Latin American Studies	3.49	3.59	3.40	3.44	3.45
Classics	3.38	3.37	3.52	3.48	3.43
German Studies	3.21	3.25	3.37	3.49	3.43
French & Francophone Studies	3.36	3.33	3.28	3.39	3.41
Mathematics/Computer Science	3.26	3.30	3.34	3.28	3.40
Physics/Astronomy	3.40	3.31	3.39	3.35	3.40
Humanities/Media/Cultural Studies	3.56	3.32	3.60	3.35	3.36
Philosophy	3.46	3.34	3.45	3.47	3.31
Linguistics Program	3.64	3.43	3.56	3.23	3.29
Geology	3.15	3.56	3.18	3.29	3.26
Biology	3.10	3.32	3.35	3.28	3.24
Economics	3.03	3.14	3.17	3.20	3.17
Chemistry	3.19	3.13	3.06	3.17	3.03
All College	**3.39**	**3.43**	**3.44**	**3.41**	**3.46**

Source: Institutional Research, Macalester College.

III. The Impact of Grading Inconsistency

As explained in the introduction, grading inconsistency is not entirely a bad thing—it represents the reality of different professors teaching different course levels and using grades to achieve different objectives. This section, however, will explore the ways in which, if left unadjusted, this inconsistency can have negative consequences. First, I will discuss the impact it has on students at the college—in selection of both major discipline as well as elective courses. Second, I will discuss the impact that inconsistency has on students when they enter the workforce as well as its impact on potential employers. Finally, I will discuss how it impacts faculty by creating pressure to continually assign higher grades.

A. Grading Inconsistency and the Undergraduate Experience

First, let's examine the impact of grading inconsistency on a student's undergraduate experience. Coupled with conversations with professors and peers, grades serve to indicate a student's aptitude for particular subject areas. Understandably, students tend to gravitate toward departments where they receive good grades. Similarly, they will drift away from departments where they receive poor grades. This is usually a healthy process—interests and abilities don't always overlap and students can adjust accordingly. However, this process becomes complicated when there is a discrepancy in how grades are applied by professors or departments.

Let's consider the hypothetical case of a student, Sue, interested in Physics, who receives a B in an introductory Physics course (while the rest of the class receives a combination of B's and C's). Meanwhile, Sue takes a leniently graded Mathematics course and is awarded an A- (while the rest of the class receives either an A or A-). After seeing the grades, Sue decides to study Mathematics. The decision could be based upon a few factors:

Sue could lack the context to understand that her Physics grade actually represents a superior performance, Sue could still feel dissatisfied with the Physics grade even after contextualization, or Sue could be fully satisfied but fear that outside observers would see the Physics grade as an inferior performance. Regardless of the motivation for the decision an unfortunate process unfolds, where Sue is driven away from a discipline (Physics) in which she have greater interest and seemingly greater talent. And, while the process is not always so simple, it unfolds on a regular basis on campuses around the country.

The point of the example stated above is not to criticize the physics professor for giving poor grades or the math professor for giving high grades. Nor is the point to criticize any department for giving higher or lower grades, on average, to students. These processes unfold naturally within any university environment. The point is that students should be given the context to understand what grades indicate about their abilities such that they can make informed decisions about their future accordingly.

Let's consider a second example. This time our hypothetical student, Dave, is a senior, pre-Med and needs to fulfill a Fine Arts requirement. For the sake of simplicity, there happen to be only two available courses which satisfy the requirement: Creative Writing and Art History. Dave is initially inclined to take the Creative Writing course, as it will likely provide much-needed improvement to his writing skills. However, after speaking to a few friends, Dave changes his mind. The Creative Writing professor has a reputation as an unusually severe grader; Dave, who had wanted to take the course precisely because he needs to improve his writing skills, is sure that he can't achieve a grade higher than B- (the average grade in the course). This poor grade, when factored into Dave's cumulative GPA, will negatively affect his chances of admission to medical school. Meanwhile, the professor of art history has a reputation as a lenient grader, where the average grade is A-. Dave takes Art

History and receives the desired grade, but there is something distressing in this example. Dave was forced to make an undesirable choice between maximizing his educational experience and maximizing his chances of admission to medical school.

This basic process unfolds on a regular basis. Students may not choose courses solely based upon grading considerations, but grades do play a part in the decision-making process. As one Macalester student explains: "I did take a lot of classes from Professor X and it certainly didn't hurt that I knew he was an easy grader. It functioned as a schedule filler…it was nice to have classes that you didn't have to put too much work into." Thus, inconsistency is advantageous in that it helps regulate workload (although workload can also be regulated by careful course selection based upon the assignments listed within the syllabus).

However, this process is troubling not only because it moves students away from course choices which would maximize their undergraduate experience, but also because of the effect of this shift in course enrollment. Severe grading relative to the norm has a two-pronged deflationary effect upon course enrollment. Students are deterred from registering for a course in the first place for fear of receiving a poor grade, while students who take the course and receive a poor grade may be deterred from taking subsequent courses in a sequence, with that particular professor, or within the department as a whole. A study by Valen Johnson (2002) at Duke University during the late 1990's found an estimated reduction in student enrollment in Math and Science courses as high as fifty percent. Unfortunately, little systematic data exists to determine the magnitude of shifts in major choices or course choices within majors. Nevertheless, the impact on elective courses alone has significant ramifications for both students and faculty.

The impact of such a decrease in enrollment on students is progressive. First, it decreases the number of students able to work in professions requiring a Math or Science background as well as decreasing the scientific competency of the general population (Neal 2006). In addition, as course enrollments fall, Math and Science departments are able to hire fewer faculty members. As fewer job opportunities become available at universities, students may be less inclined to pursue Math and Science (at least at the Ph.D. level). One risk of such a decrease in Math and Science education is that it could affect national scientific competitiveness, which is critical because: "In today's global economy, the ability of the United States to remain competitive relies increasingly on our ability to develop and commercialize innovative technologies." (Jeffrey 2006)

B. Grading Inconsistency and the Post-Graduate Experience

Grading inconsistency also negatively affects both employers and students, particularly traditionally disadvantaged students, because it undermines the potential of grades to serve as accurate evaluative tools. While grades are far from perfect measures of academic performance, they are preferable to the alternatives: test scores and networking.

> When all students receive high marks, graduate schools and business recruiters simply start ignoring the grades. That leads the graduate schools to rely more on entrance tests. It prompts corporate recruiters to depend on a "good old boy/girl" network in an effort to unearth the difference between who looks good on paper and who is actually good.
>
> Put to disadvantage in that system are students who traditionally don't test as well or lack connections. In many cases, those are the poor and minority students who are the first in their families to graduate from college. No matter how hard they work, their A's look ordinary. (USA Today 2002)

As grading inconsistency forces an increased reliance on test scores and connections, traditionally disadvantaged groups of students will lose the opportunity to demonstrate their

ability through receiving exemplary grades. Employers, on the other hand, will end up settling for potentially less qualified employees, since applicants' college performance is obscured by grading inconsistency.

C. The Impact of Grading Inconsistency on Professors

In assigning grades, professors are traditionally compelled to reconcile two competing goals. While lenient grades will serve to attract a greater number of students, severe grades will encourage those students who do enroll to take the course seriously and complete coursework in a thorough and timely manner. Some professors navigate this conflict by only using grades in pursuit of only one of these ends; however, this strategy can put them at a disadvantage. If they grade leniently to attract students, they are robbed of an effective motivating tool. Instead of leveraging a student's grade, these professors are forced to fall back on more informal motivators: coaxing and empty threats. Professors who grade with sufficient severity to motivate students, on the other hand, put themselves at a competitive disadvantage relative to their peers in attracting students. Even if the severely grading professor offers an attractive syllabus and convenient course times, other professors still offer the comparative advantage of more lenient grading. This competitive incentive can produce a "race to the bottom" where coursework expectations continue to fall and grades continue to rise as professors strive to keep pace with their peers. It should be noted that everyone is worse off in this case: grades no longer act as an enrollment incentive if all professors give similarly high grades, nor are they able to function as a motivating tool since students can now predict that they will receive high grades regardless of performance. The scenario has the characteristics of a classic collective action problem. No individual

professor is at fault, and individual decisions to offer better grades are rational. An effective solution can therefore only come from action by the college as a whole.

Stringent grading also has been shown to negatively affect the course evaluations received by professors. In a study at Duke University, Valen Johnson (2003) showed that there was a strong connection between grading practices and the course evaluations received by professors.

> Student responses to the survey were significantly affected by the grades that the students either expected to receive or already had received. For most items, the influence that students' grades had on their responses to the survey ranged from about one-fourth to one-half the importance of the consensus rating variable estimated from the responses collected from all students who took the course. This suggests that although the consensus opinion of instructional attributes was the most important predictor of students' responses to an item, grades do, in fact, represent a serious bias to student evaluations of teaching. (Johnson 2003, 100)

Johnson suggested that the bulk of the change in evaluations stemming from grades could be explained by "grade attribution theory", where students associate bad grades with bad teaching and associate good grades with good teaching (Johnson 2003, 100). Johnson's study is hardly the first to examine the connection between grades and course evaluations. However, it is relatively unique in that it used an experimental instead of an observational format[2]: course evaluations were given to students before and after the distribution of final grades such that the main explanation for changes in perception would have to be related to the grades received by students.[3]

The cumulative impact of grading inconsistency on faculty who grade with relative stringency is that they receive both poor course evaluations and lower course enrollments.

---

[2] The handful of other experimental studies, such as those conducted by Holmes (1972), Vasta and Sarmiento (1979), and Chacko (1983), produced similar results.

[3] For a more complete survey of research examining the connection between grades and course evaluations, see Chapter 3 of Valen Johnson's book *Grade Inflation* (Johnson 2003).

Thus, these professors are "less likely to receive tenure, salary increases, and promotions" (Johnson 2003, 9) not because they teach poorly but simply because they approach grading differently than their colleagues.

IV. Taking Action

Each of these factors taken in isolation—the impact of grading inconsistency on undergraduate experience, on post-graduate opportunities and the negative impact on faculty—would warrant action to remedy the issue; the combination of all three makes the value of taking action even greater.

The most common proposal is that the change should come from the professors themselves or the departments. This proposal is both untenable and ill-advised. It is untenable because the process of coordinating grading practices among dozens of professors or a number of different departments would be a logistical nightmare unless a strict curve with a set average was required for all courses. It is ill-advised because it interferes with individual professors, who ought to have basic autonomy in crafting their approach to teaching.

Any effective strategy to remedy grading inconsistency must be implemented by the college or university at the institutional level. As the central repository for grade information, the college possesses the ability to provide sufficient information to contextualize grades such that performance is measured appropriately. This adjustment does not endorse the use of grades to measure academic performance. Instead it recognizes that grades *are* currently used as measures of academic performance and, given this reality, seeks to minimize the potentially deceptive effects of grading inconsistency.

This is not entirely uncharted territory; a few pioneering schools have undertaken efforts to alter grading practices. In the subsequent sections, schools which have implemented a particular grading reform are noted and their success in altering grading behavior is evaluated.

The review will proceed as follows. First, I will examine proposals which target only grade inflation. Next, I will examine proposals which seek to address both grade inflation and grading inconsistency. Finally, I will examine proposals only concerned with addressing grading inconsistency. Within each section there is no precise ordering, although I will generally try to examine policies in increasing order of sophistication (or complication, depending on your perspective).

A. Anti-Grade Inflation Policies

i. Inclusion of Class Rank

This past fall (of 2006), the University of Colorado became the latest university to publicly hash out the grade inflation debate. University President Hank Brown took the lead in a charge to change the college's policies to contain grade inflation. He argued that the college should include either a class ranking or grade percentile (within the student body) on the transcript so that employers and graduate schools could accurately contextualize a student's GPA (Brown 2006a). After a lively debate, where the proposal received largely positive treatment from the media (most notably from Kurtz 2006), the Board of the college opted for a compromise: class rank would be disclosed with transcripts, but only at the request of students (Brown 2006c). The compromise was reached following "spirited protest from professors who opposed tougher, mandatory remedies to curb grade inflation." (Brown 2006c)

Colorado is only the most recent case where political pressure from faculty has foreclosed the possibility of more fundamental changes to grading practices. This is not to say that lively debate from different perspectives is bad, but solutions such as the one adopted by Colorado have more symbolic value than actual impact. Including class rank on a student's transcript upon request is a positive step but a small one. Class rank does allow outsiders to evaluate the relative performance of a student better than GPA alone. However, it is unlikely to do much to address grade inflation. Students would have to decide that it is in their interest to pressure professors to distribute lower grades and professors would have to respond accordingly—a relatively implausible scenario. More problematically, class rank does nothing to adjust for grading inconsistency within a graduating class and, thus, can give an incorrect impression of the relative standing of students within the institution.

ii. Restricting Students Who Graduate With Honors

> Some say Harvard students are better these days and deserve higher grades. But if they are in some measures better, the proper response is to raise our standards and demand more of our students. Cars are better-made now than they used to be. So when buying a car, would you be satisfied with one that was as good as they used to be? (Mansfield 2001)

Nowhere has the debate over grade inflation been more intense or more highly publicized than at the Ivy League schools. Harvard, in particular, came under scrutiny in 2001 when it was revealed that 91% of students had graduated with some sort of honors (Healy 2001). While Harvard does not necessarily give substantially higher grades, on average, than its peer schools, it was unique in that it combined such high GPA's with lower GPA thresholds for receiving honors. Peer schools, such as Yale and Princeton, granted honors to approximately half as many students (Healy 2001).

The resulting controversy produced two grade-related reforms at Harvard. The first reform was to abolish Harvard's idiosyncratic 15-point grading system which weighted the gap between an A- (14 points) and a B+ (12 points) twice as much as the gap between an A (15 points) and an A- (Healy 2002). The idea was to encourage more professors to make the downward adjustment to a B+ (from an A-) as the gap would no longer be as large (Crenshaw 2002). The second reform was to cap the number of honors degrees while encouraging professors to give fewer A's. Under the new policy only 60% of students would be eligible to receive honors, and professors would be notified if their grading standards differed unreasonably from those of other professors at the college (Meyer 2005).

These reforms appear to have done little to impact grading practices at Harvard. Students still receive A's at approximately the same rate (Boston Globe 2004) and mean GPA for graduating students has decreased only from 3.42 to 3.41 (Meyer 2005).

iii. Deemphasizing grading

Some colleges have managed to control grade inflation without resorting to any particular policy remedy, by establishing a culture conducive to more severe grading. The most notable example is Reed College, a selective liberal arts college in Oregon, where the mean GPA is a relatively meager 2.90 (Neal 2006). At Reed, grades are largely removed from the learning process.

> Papers and exams are generally returned to students with lengthy comments but without grades affixed. (Reed 2007)

Additionally, students do not receive their grades at the end of each semester as at most colleges; instead they only receive their grades upon request.

While Reed is an exemplar in controlling grade inflation, it also demonstrates the danger of such a policy.

> Colin S. Diver, Reed's president, says graduate schools worried about their rankings are becoming less willing to take students with lower grades because they make the graduate schools appear less selective.
>
> "If they admit someone with a 3.0 from Reed who is in the upper half of the class, that counts against them, even if it is a terrific student," Mr. Diver said. (Arenson 2004)

The argument made by Diver is a refrain among those who oppose efforts to end grade inflation. While it may be a good idea in theory to adjust for grade inflation, the argument goes, it can't be done without hurting the students of that particular university. Given this concern, efforts which put greater emphasis on grading inconsistency rather than grade inflation may be more prudent.

B. Policies Addressing Both Grading Inconsistency and Grade Inflation

Not all efforts to address grading practices have been limited to grade inflation measures. A few daring schools have adopted measures far stronger than those pursued by Harvard and Colorado.

i. Median Grades

A commonly proposed strategy, which few schools have chosen to pursue, is to include median grades for each course next to a student's grade so that an outside observer can gauge the relative difficulty of each of the student's courses. One school, Dartmouth, has included both the median grade for each course as well as the course size on all student transcripts since 1994. Unfortunately, the proposal has been unsuccessful in controlling

grade inflation; the average grade point has actually risen at Dartmouth over that span from 3.25 to 3.33 (Gardner 2002).

While the inclusion of additional information could be helpful in contextualizing a student's grades, this method has two critical shortcomings as a correction for grading inconsistency. First, it fails to provide any indication of the strength of other students in a course. A grade equivalent to the median grade in a course with all top students would be quite an impressive performance, while receiving the median grade in a course with weaker students represents a notably lower performance. Certainly it is possible to make some estimate of average student strength according to course level and content, but this risks providing a misleading picture of student achievement if the composition of a course is at all counter-intuitive. Second, the process of trying to contextualize over thirty different grades is unwieldy. For each course, the evaluator is required to make a somewhat complex and imprecise approximation of the "true" achievement that a particular grade represents. In practice, then, the inclusion of median grade and course size is likely to do little to aid those on the outside in understanding a student's grades other than to create confusion.

ii. 'A' Quotas

In 2004, professors at Princeton voted to impose a quota on the number of A's given at the college, setting a loose cap at 35% of all grades (Brown 2006b). The initiative has been at least somewhat successful, as the percentage of A's awarded fell from 46% to 41% in the first year after the initiative (Mount 2005). Moreover, while the proposal on face only targets grade inflation, it will likely impact grading inconsistency as well. Professors are being pushed to give a more uniform percentage of A's in a each course, meaning that professors will be forced to assign grades in a more consistent manner. Additionally, this

policy decreases grading inconsistency produced by the assignment of large quantities of A's, which fail to distinguish between excellent and more average performances.

However, the policy still appears to have its limitations. First, despite trending in the right direction, the number of A's awarded at Princeton still exceeded the cap by 6% in 2005, meaning that the cap functions more as a guideline than a regulation; a fact which could undermine its effect. Second, by not considering the strength of students enrolling in a particular course, the quota has the potential to penalize students enrolling in difficult courses with other strong students. The more effectively the cap is enforced, the more it creates a disincentive for students to take valuable, upper-level courses in departments.

C. Mathematical Corrections for Grading Inconsistency

Finally, there are proposals for more precise mathematical adjustments of grades to determine a more accurate adjusted GPA or class rank. Such adjustments have rarely made it past the proposal stage as they usually face substantial opposition. The most notable example is the Duke case, where the grading reform proposal failed by just a handful of votes, due in part to the complicated nature of the mathematical adjustment. These proposals do not purport to address grading practices themselves (i.e. they don't directly affect grade inflation) but instead are designed to produce alternate, more accurate, GPAs or class ranks.

i. Linear Adjustment

The most common proposal for mathematically adjusting grades is to linearly adjust the grade according to the difficulty of the course. The most basic linear adjustment is based upon research done by Goldman and Widawski (1976) and Strenta and Elliot (1987), where they found that differences in departmental grading practices were additive (i.e. if Physics courses are, on average, 0.2 units of GPA harder than Biology courses and Biology courses

are 0.2 units harder than Math courses, then Physics courses will be 0.4 units harder than Math courses).

The proposal of Elliot and Strenta (1988), then, was to determine the additive component for each individual course. The adjusted grade would thus be:

Formula 1: Additive adjustment to grades as per Elliot and Strenta (1988)

$$\text{Adjusted Grade} = \text{Nominal Grade} + \beta$$

The constant (β) in the above formula represents course difficulty—a positive constant indicates a relatively difficult course, while a negative constant indicates a more leniently graded course. The additive constant was determined by averaging the pair-wise differentials between a department and each of the other departments.

Their proposal has two advantages. First, it does provide a somewhat accurate correction for grade inflation. Students are not penalized for taking courses with more stringently grading professors; nor are they penalized for taking courses with stronger students. Second, it is simple. The concept of adjusting grades upward or downward a bit based upon course difficulty is intuitive and the fitting process requires only an understanding of basic algebra.

The proposal, however, does have its drawbacks. First, it would penalize top students for taking more leniently graded courses. If a course had an additive factor of -0.5, a student could not achieve better than a 3.5 in the course, no matter how well they performed. Second, the method fails to account for the manner in which grades are distributed. For example, let's take two courses with the same nine students and both courses have a median grade of C. In one course, the grades are distributed such that there are 3 A's, 1 B, 1 C, and 4 D's (the grades are "spread"). In the other course, there is 1 B, 7 C's and 1 D (the grades are "bunched"). The grade of B would be treated identically by the

linear adjustment method despite its greatly different meaning. In the first course with spread grades, a B appears to represent approximately an average performance. Meanwhile, in the second course with bunched grades, a B represents exemplary performance—superior to that of the other eight students.

A more sophisticated fitting procedure is proposed by Caulkin, Larkey and Wei (1996) of Carnegie Mellon University. Their linear adjustment fits a line to student ability (θ), translating ability to a predicted grade for each course:

Formula 2: Linear adjustment to GPA as per Caulkin, Larkey, and Wei (1996)

$$\text{Predicted Grade} = \alpha * \theta + \beta$$

The fit is done by minimizing the squared error between this predicted grade and the actual grade that the student received. Because both ability of the student and the course difficulty parameters are unknown, the authors do the fit iteratively. To begin this iterative process, one of the parameters must be assumed—the easiest assumption would be to set ability equal to the student's raw GPA.

This method represents a notable improvement upon the simple additive procedure previously used. Most importantly, it corrects for differing distributions of grades. The above example (which consisted of a course with nine students) showed that the additive method did not do well in comparing a course with quite spread grades and a course with tightly packed grades. However, the α coefficient within this new linear adjustment formula addresses the problem. A small α coefficient deemphasizes student ability in predicting grade allowing for a tighter packing of grades. A large α coefficient magnifies differences in underlying student abilities producing a greater spread in the distribution of grades.

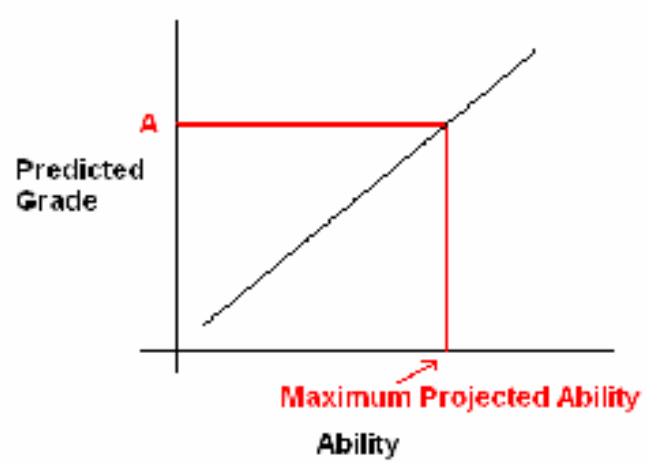
Still, the method has its shortcomings. First, it relies on the strong assumption that professors treat the difference between an A and A- the same as between an A- and B+.

Undoubtedly, this is true in some cases. However, there are cases where the professor takes a more idiosyncratic approach to grading. Let's consider a hypothetical professor who only assigns A's in truly exceptional cases (when performance far exceeds that of an A-), but considers the difference between a B+ and B to be rather trivial. The linear adjustment method would treat this distinctions as equivalent when, in fact, they are not.

A second, more important, shortcoming is that the method would still penalize top students for taking courses with weaker students. Let's take a hypothetical case where a top Mathematics student, Jane, takes a series of difficult courses with top students in the department and achieves outstanding grades in all of them, producing a high student ability parameter. Jane also has a passion for Art and takes a number of courses in the Art department, completing an Art minor. Her work in the department is exemplary—she achieves A's in all of her Art courses. The students in Jane's Art courses have substantially lower ability parameters but still achieve A's in the same courses she took. In fitting the line, Jane's ability parameter will be dragged down to match more closely to those other students who received A's in Art courses. Thus, while nothing about her performance in the Art courses indicated that Jane is a weaker student, her ability parameter (θ) falls simply because she enrolled in the course. This phenomenon is shown in Figure 2.

This concern may be relevant to more than just a few top students; there would be a disincentive for any above average student to take courses in departments where students generally tend to have weaker ability profiles. These students, like Jane, would be forced to make the artificial choice between maximizing their ability parameter (and corresponding adjusted GPA or class rank) or taking the courses from which they believe they will derive the greatest educational benefit.

Figure 2: Linear adjustment graph showing potential penalty to top students taking leniently graded courses.



ii. Real vs. Nominal GPA

A proposal by Felton and Koper (2004) is to calculate what they call a "Real GPA". Their proposal focuses on the importance of a student's standing relative to mean GPA of students taking the course (referred to as "Class GPA" within the paper). The formula they use to calculate this Real GPA is quite simple:

Formula 3: Calculation of Real GPA, as per Felton and Koper (2004)

$$\text{Real GPA} = \sum 2 * \left(\frac{StudentGrade}{ClassGPA} \right)$$

The result is a series of student grades for courses which center around 2. The proposal would certainly curb inflation in grades—it would be impossible to give grades which, on average, exceed two points of real GPA for any individual course. Moreover, it would take a step toward addressing grading inconsistency—all grades would be reduced to a common scale.

However, this adjustment for grading inconsistency overlooks one crucial factor: variation in student strength by course. While sometimes this variation is minimal, there are situations where it can be notable. Consider the case of a senior deciding between an upper-

level elective and an introductory elective. One is likely to have students with three more years of schooling and have far greater background in the subject area, while the other is with younger students who may know little or nothing about the subject area prior to the course. Yet, Real GPA would only ask how the student ranks relative to other students in the course and would reward that student for taking the introductory course.

Such a method could, in fact, introduce grading inconsistency where none existed previously. Let's say that a professor is teaching two Economics courses: Microeconomics and Macroeconomics. Throughout the semester, the professor determines that the students in her Microeconomics course are superior—they turn in higher quality work in a timelier manner and are far more informed during class discussions. Accordingly, the professor assigns grades at the end of the semester such that the Microeconomics course has a mean Class GPA 0.5 higher. In assigning these grades, the professor is conveying information about the relative quality of students within the course. Yet this information is lost through calculation of Real GPA.

iii. The Achievement Index

Around ten years ago, a Mathematics professor at Duke, Valen Johnson, developed what he terms the "Achievement Index". Following a controversial debate on the implementation of the index, Johnson's method was rejected mainly because it was deemed to be too complicated.

The index begins with basic assumption that professors are able to accurately order students. After that, no assumption is made about the particular shape of the grade distributions. Cutoffs for grade levels within each individual course and student ability are

determined using Bayesian fitting techniques. The mathematics behind this fitting are omitted to avoid unnecessary confusion.[4]

The Achievement Index does avoid the aforementioned shortcomings of the other adjustment techniques. It corrects for inconsistency in grade distributions, and the quality of students in each course, such that there is not an excessive penalty for taking courses with weaker students (or an excessive reward for taking courses with stronger students). Nevertheless, the complicated nature of the technique limits its utility within a university setting where the comprehensibility is likely a prerequisite for implementation.

iv. Other Potential Adjustment Methods

There are other adjustment methods which have not yet been applied to grade data. American college hockey ranks teams according to a maximum likelihood process where the probability of a team with ability x defeating a team with ability y is simply[5]:

Formula 4: Probability of a victory according to the American college hockey ranking system

$$P_{win} = \frac{x}{x + y}$$

This method could be applied to grades by constructing a series of pair-wise "competitions" between students in a course. One would treat one student achieving a better grade than another as having scored a "win" over that student in the course. The biggest drawback to the method is that it struggles to account for unbeatens within the system (in the academic case this would be 4.0 students) as their ability parameter would be infinite. There is also the difficult question of how the system would deal with ties (i.e. when two students receive the same grade in class).

[4] For further discussion of Johnson's fitting technique, see Johnson (1997) and Johnson (2003).
[5] The adjustment method is termed KRACH after its creator. For more information on the technique, see the American College Hockey's website: http://www.uscho.com/rankings/?data=krach.

Another potential avenue for grading adjustment methods would be to consider multiple academic ability parameters. Berry (2006) models professional football using separate offense and defense parameters in order to more accurately predict game outcomes. Similarly, grades could be predicted by separating different types of academic abilities. A possible division could be Math and Verbal (like the SAT) or the degree to which the course is paper or exam based. This unexplored avenue could be incorporated within the aforementioned linear adjustment procedure as well as Item Response Theory, which will be discussed in the following section.

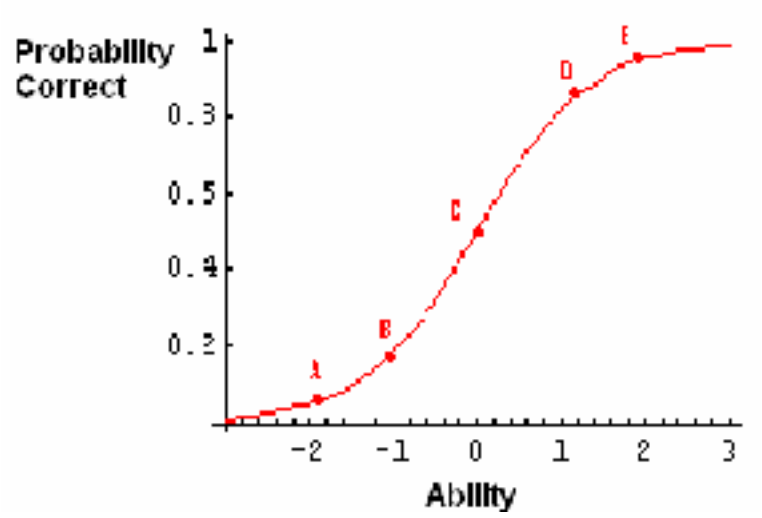
V. Item Response Theory

A. Background

Item response theory is designed to measure latent traits—abilities which aren't measured directly, but only indirectly through the application of tests designed to draw upon that trait. Item response theory is designed not just to measure these latent traits but also to test their consistency between items.

A typical item response curve looks like this:

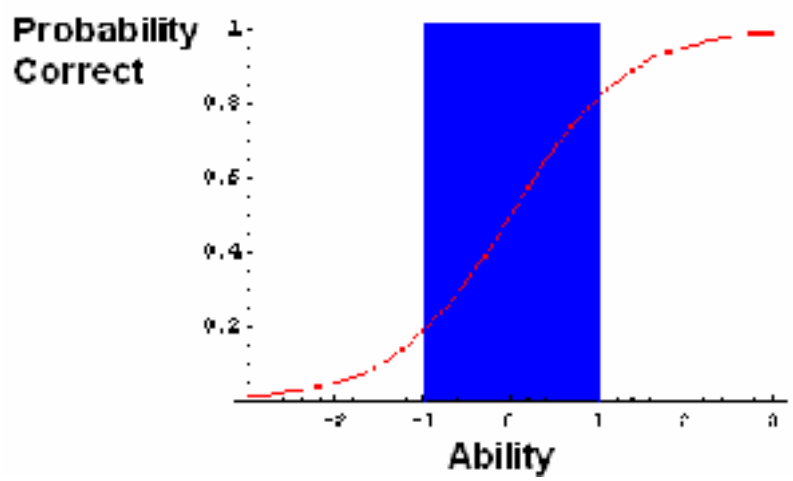
Figure 3: Typical Item Response Curve



The curve measures the probability that an individual at a given ability level will correctly respond to a dichotomously scored question (meaning that there is no "partial credit", an answer is either right or wrong). Those at low ability levels (points A and B) are unlikely to answer the question correctly. Those at high ability levels (points D and E) are quite likely to answer the question correctly. An individual at a middle ability level (point C) has about a fifty percent chance of answering the question correctly.

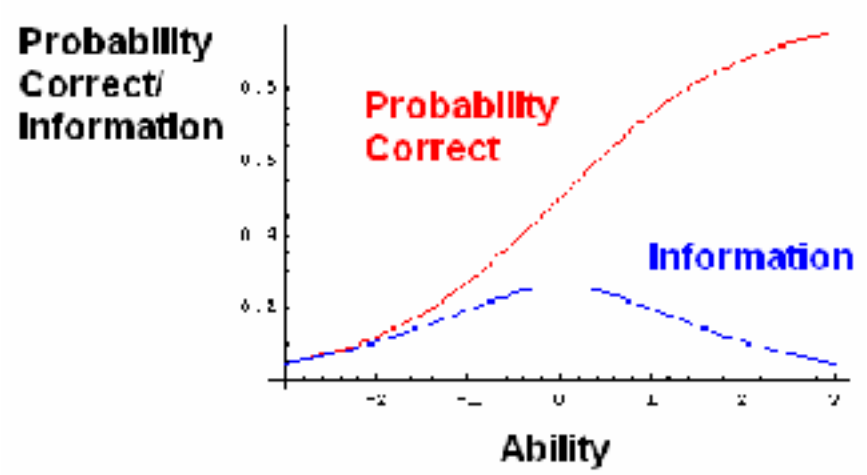
Looking at the curve, it appears unlikely that we will learn anything which would allow us to differentiate abilities A and B or abilities D and E. Instead, the bulk of what we learn is concentrated in the middle section between points B and D. In item response theory, the term "information" is used to refer to how much a given item (or test question) tells us about an individual with a given ability. Figure 4 highlights the ability levels which the question tells us the most about.

Figure 4: Sample IRT curve showing ability ranges where maximum information is obtained about students



The information function can be represented more precisely as a continuous function over the same range as the item response curve. An example is shown in Figure 5:

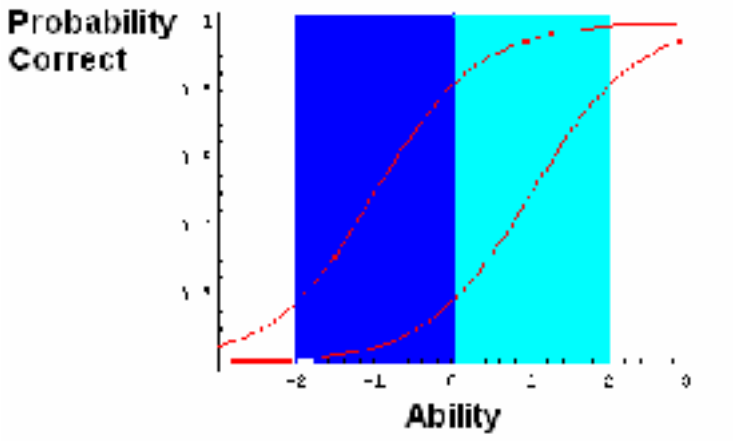
Figure 5: Sample item response curve shown with the corresponding information curve



Notice that the information function reaches its maximum at the inflection point of the item response curve. This convenient fact will make it easier to predict subsequent behavior of the information function.

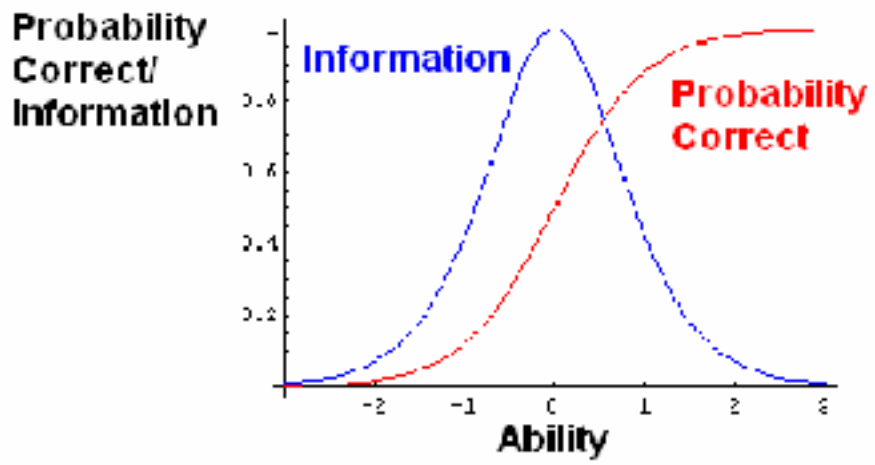
A good test, of course, will offer information about students at a variety of ability levels. If all questions targeted those at middle ability levels, then we would know a great deal about those students but would lack the ability to differentiate the best students from the above average students and to differentiate the worst students from the below average students. The following graph displays two item response curves which provide information on different levels of subjects:

Figure 6: Two item response curves which provide information on students at different ability levels

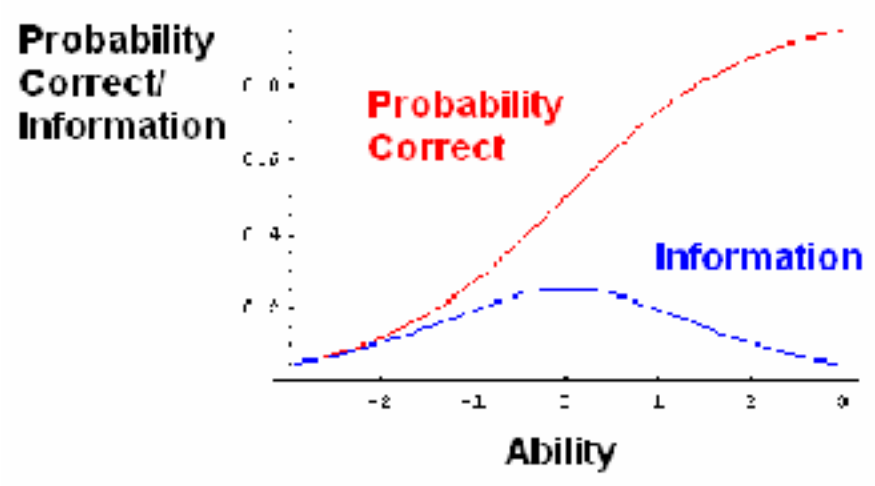


Of course, not all item response curves have the same shape. As the graph above indicates, curves can be centered anywhere along the ability axis, as well as possessing different slopes. The curve shown in Figure 5, for example, is also far steeper than the curve shown in Figure 6. Steeper curves offer information about individuals within a narrow band of ability but provide more information about these individuals. Shallower curves offer information about individuals within a broader ability range but provide less information about them. The difference in the shape of the information functions can be seen in Figures 7 and 8.

Figure 7: An item response curve with a steeper information function.



Figures 8: An item response curve with a flatter information function.

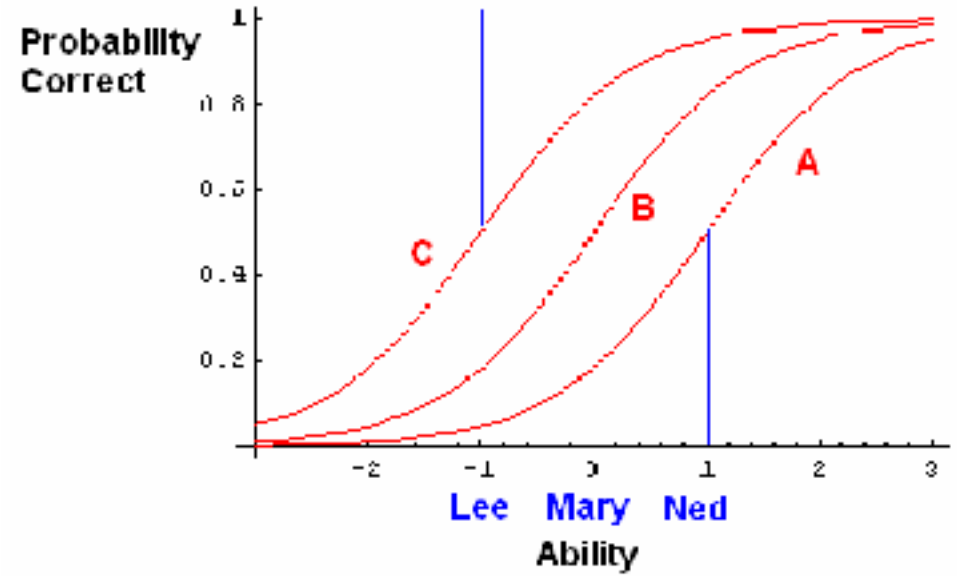


One last convenient feature of information curves is that they can be summed between items. This allows for the use of individual items to target different levels of ability, which taken together allow the entirety of the test to distinguish between individuals at all levels.

B. Application of IRT to Grades

The application of IRT to grades is not entirely obvious since they are not dichotomous outcomes. However, whether or not a student successfully reaches a particular grade threshold can be treated as a dichotomous outcome. John Young, the one scholar who has applied item response theory to grades (Young 1989; Young 1990), uses A, B, and C as the grade cutoffs. There is nothing about the procedure which restricts one to consideration of those particular grade cutoffs—those were simply the cutoffs which Young found to be most convenient. A sample graph is shown below with students at three different ability levels.

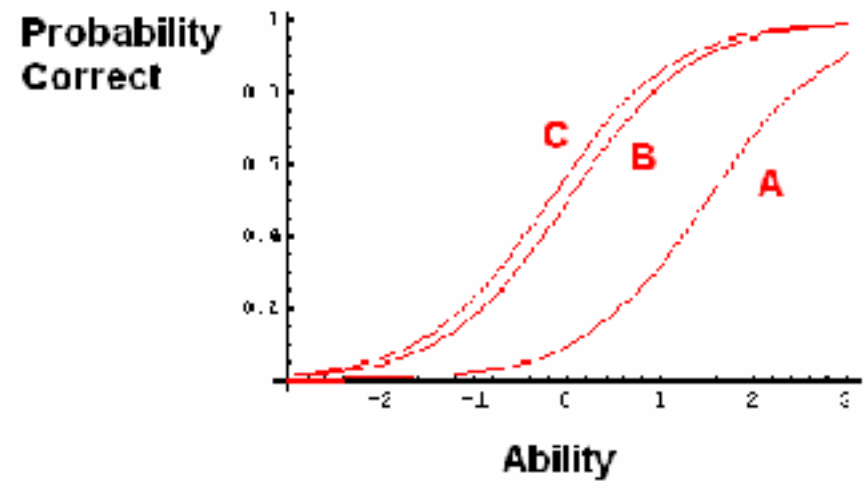
Figure 9: Sample IRT curves for grades of A, B and C with three sample students shown



Looking at the first student (Lee), we see that Lee is unlikely to achieve either an A or B within the course (the curves cross the ability line at a very low probability). However, Lee is more likely to achieve at least a C (the ability line crosses the item response curve for a grade of C at a higher probability. Thus, we would predict that Lee would receive a C in the course. Using similar logic, we would expect Mary to receive a B, and Ned to receive either an A or B.

The application of IRT to each individual class produces a set of item response curves which provides information about a range of student abilities. However, not all courses will have this convenient distribution where the curves are spaced to maximize information about students. The following graph shows what a set of curves would look like if a professor refused to give C's except under exceptional circumstances (by jumping straight from B to D in assigning grades).

Figure 10: IRT curve for a sample course showing a bunching of the item response curves for the grades of B and C, indicating a reluctance of the professor to assign the grade of C

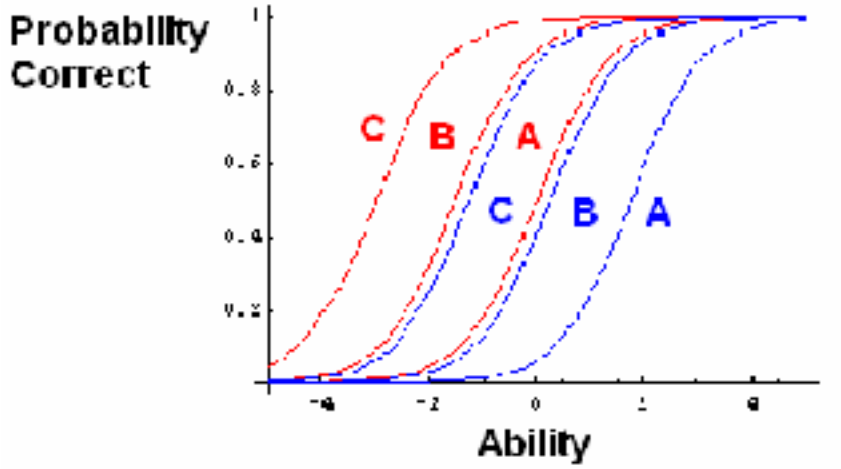


The proximity of the B and C curves indicates that an individual student will have approximately the same probability of receiving a B (or C) or better. In order for there to be

C's distributed to students, we must observe a gap between the B and C curves and this only occurs at a small sliver of ability levels.

The necessity of correcting for grading inconsistency can be seen from the following graph where two classes are compared (one in blue and one in red), where a B in the blue class represents a better performance than an A in the red class, and a C in the blue class represents a better performance than a B in the red class.

Figure 11: Item response curves demonstrating the importance of adjusting for grading inconsistency



C. Mathematics Underlying IRT

Thus far I've only spoken about the basic graphical qualities of IRT. Now I'd like to introduce the underlying mathematics. The basic formula for an IRT curve has three variables[6]:

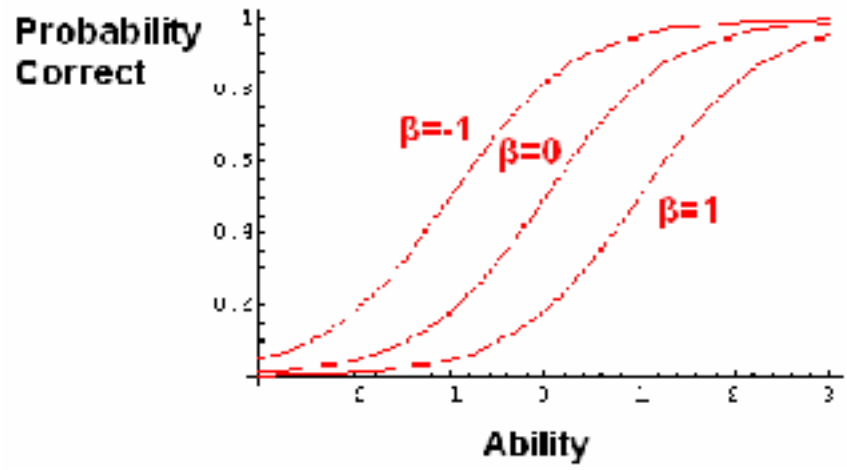
Formula 5: Expression for the basic item response curve

$$P_{correct} = \frac{1}{1 + e^{-\alpha(\theta - \beta)}}$$

[6] It should be noted that although the item response function is described in this section, the fitting technique is not. This discussion is omitted to avoid unnecessary distraction. All fitting techniques for IRT use some sort of maximum likelihood estimation. The best fitting techniques are marginal maximum likelihood estimation (MMLE) and the EM (ExpectationMaximization) algorithm. For more information on these two approaches, see Baker (1992).

θ is the parameter which estimates the ability of an individual student. β is known as the difficulty parameter and is used to determine where the curve is centered.

Figure 12: An item response curve showing the horizontal shift with changes in the value of β.



When θ and β are equal the probability that a student will answer the question correctly (or reach a particular grade threshold) is one-half:

Calculation 1: Showing how β, the difficulty parameter, functions

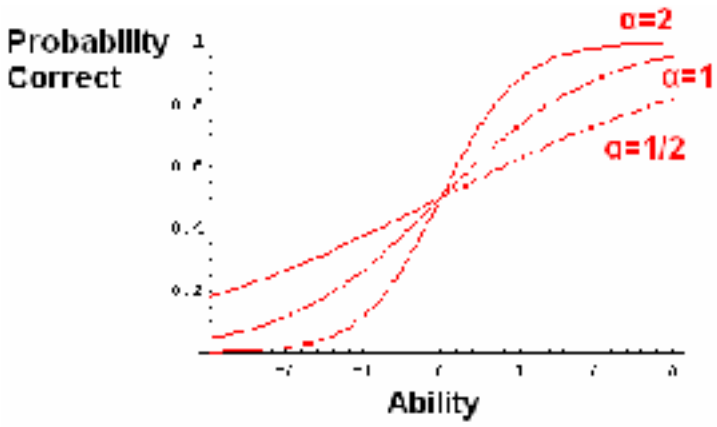
$$P_{correct} = \frac{1}{1+e^{-\alpha(\theta-\beta)}} \text{ and given that } \theta - \beta = 0$$

$$P_{correct} = \frac{1}{1+e^{0}} = 1/2$$

The ability level where θ is equal to β is also where the inflection point occurs in the item response curve and, thus, the curve is at its steepest slope. This is the point at which the item response curve provides the greatest amount of information about the ability of the student.

The α parameter is known as the discrimination parameter. α determines the slope of the curve—a high value of α will produce a steeper item response curve and a lower value of α will produce a shallower curve. The following figure shows how this α value affects the shape of the item response curve:

Figure 13: Sample item response curves showing change in steepness as the value of α varies



Because the α parameter determines the slope of the item response curve, it correspondingly determines the shape of the information function. As explained earlier, a shallower curve, which is produced by a lower α value, will produce a wider, flatter information curve. A steeper curve, produced by a larger α value, will produce a narrower, steeper information curve[7].

There are a few variations of the previously described two parameter logistic curve[8]. The first variation is a one parameter model known as the Rasch model. The Rasch model holds the discrimination parameter (α) constant and only fits the curve to the difficulty parameter (β).

The second variation is the three parameter model which adds an additional "guessing" parameter to the two parameter model. The "guessing" parameter, c, is designed for multiple choice tests where the probability of getting the correct response asymptotically approaches c instead of zero. For example, in a multiple choice test with four possible responses, an individual's probability of correctly answering should approach 0.25, as shown

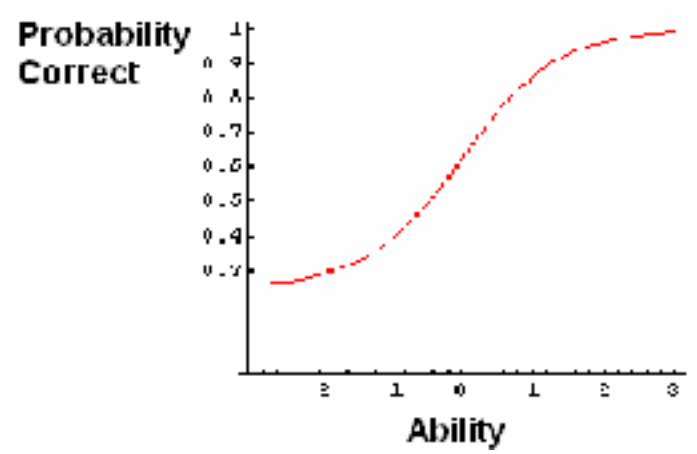
---

[7] The formula for the information curve is: $I(\theta) = \alpha^2 * \left(\frac{1}{1+e^{-\alpha(\theta-\beta)}}\right) * \left(1 - \frac{1}{1+e^{-\alpha(\theta-\beta)}}\right)$

[8] It is referred to as a two parameter model because only the α and β parameters determine the shape of the curve. The third variable, ability (θ), is the independent variable and, thus, is not considered a parameter in the construction of the model.

in Figure 12 (unless answer responses successfully trick respondents into choosing wrong answers with higher frequency).

Figure 14: Sample three parameter IRT curve with a "guessing parameter" equal to 0.25



The inclusion of this third parameter impacts some of the aforementioned convenient features of the item response curve. First, the inflection point still occurs when ability (θ) is equal to the difficulty parameter (β), but it occurs at a probability of (1-c)/2 instead of at 0.5. Second, it complicates the calculation of the information function.

Neither model will be used in this paper. The Rasch model's failure to account for variations in item discrimination makes it overly simplistic. The three parameter model on the other hand introduces an unnecessary parameter—with grades there is no equivalent to guessing, if a student does particularly poor work, or doesn't even do the work, then they will fail to achieve all grade levels except F (which doesn't constitute achieving a grade level but represents a failure to achieve any passing grade level). The three parameter model also undermines some of the convenient mathematical properties of the two parameter model, complicating the calculation of both the item response curve and the information function.

The final variation is that some authors use a normal ogive curve (which has the same shape as a cumulative normal distribution) instead of a logistic curve (Lord and Novick

1968). The fitting of the curve is not substantially different—indeed if adjusted by a multiplicative factor, the normal ogive will produce almost an identical fit to that of a logistic curve (Baker 1992). The mathematical convenience of a logistic function, however, makes it preferable for use in fitting grade data.

D. Comparative Advantages of IRT

IRT has a few comparative advantages as a grade adjustment technique. First, and most importantly, it corrects for irregular distributions of grades (when grades are either skewed in some way or irregularly clumped). By treating each grade level as a dichotomous outcome with an independently determined distribution, no problematic assumptions are made about the relative difference between an A and A-, and a A- and B+. Other adjustment methods, such as linear adjustment, assume that the distance between grade intervals is uniform.

Second, IRT never penalizes students for receiving an A in a course. While some adjustment methods penalize top students for taking easier courses and receiving an A, IRT ensures that a top grade can only benefit a student—only when they fail to achieve an A can their ability parameter fall. Thus, while a student will benefit more from receiving an A in a more severely graded course, there is no disincentive to taking a more leniently graded course if the student believes that the course would be beneficial educationally. I should note that this argument relies on the minimax criteria (minimizing the maximum error of all student ability estimates) for model selection as opposed to the traditional least squares approach. This is because it is the cases of bias where pressures are created to alter course selection, such as that shown in Figure 2.

The study by Caulkin, Larkey, and Wei (1996) found that the linear adjustment notably outperformed IRT using least squares as the criteria (the R^2 of adjusted GPA relative to exogenous measures of student ability was 0.321 for linear adjustment and 0.264 for IRT). Besides not being the appropriate criteria for comparing adjustment methods, I believe there are two shortcomings within their study. First, it used exogenous measures of student performance, which are subject to the limitations discussed earlier. The low R^2 value for each of the adjustment techniques demonstrates these shortcomings. Second, their implementation of IRT relied upon the methodology outlined in Young (1989), which uses grade cutoffs of A, B and C. By only using these cutoffs, valuable information on student performance is lost. No distinction is made between an A- and a B+ or a B+ and a B. In practice, this is where professors do the bulk of their sorting among students, as shown in the following table which shows the distribution of grades received by Macalester students in all courses.

Table 5: Grade distribution for all students at Macalester, Fall 2001-Fall 2005

Grade	**Fall 2001**	**Fall 2002**	**Fall 2003**	**Fall 2004**	**Fall 2005**
A	24%	28%	26%	26%	28%
A-	20	21	22	23	25
B+	18	17	17	18	17
B	14	14	14	14	13
B-	7	5	6	6	5
C+	3	3	3	2	2
C	3	3	3	2	2
C-	1	1	1	1	1
D+	*	*	*	*	*
D	1	*	*	*	1
D-	*	*	*	*	*
NC	1	1	1	1	1
I	2	1	1	2	1
SC	6	6	6	5	5
GPA	**3.34**	**3.39**	**3.39**	**3.40**	**3.43**

Source: Institutional Research, Macalester College.

Third, IRT produces an information function which can serve as a useful resource for faculty. If, at the end of each semester, faculty members were provided with a report on their grading which includes the information function for each of their courses, they would be able to see which students they are sorting well and which students they are sorting poorly. A professor could then choose to adjust grading practices accordingly.

Fourth, IRT is potentially more acceptable than other adjustment techniques involving complicated mathematics, such as the Achievement Index, for two reasons. First, the item response curve behaves intuitively (the probability of a student answering a question correctly, or achieving a particular grade level, increases as their ability parameter increases) and this behavior can be shown graphically. Accordingly, the method can be explained and (more or less) comprehended without a complete understanding of the underlying mathematics. Second, IRT is currently used to calibrate standardized tests such as the SAT. While few people are aware of this, I believe that this use of IRT can be cited as evidence of the method's efficacy as well as its acceptance within a portion of the academic community. Thus, although the Achievement Index shares many of the advantages of IRT, I believe that IRT is a superior method.

VI. Simulation to Show Efficacy of Grade Adjustment Techniques

Although the previous section provided arguments as to why item response theory is the *best* method to adjust for grading inconsistency, this is not to imply that the other adjustment techniques surveyed earlier in the paper are bad techniques. Indeed, a number of those techniques are at least moderately effective at adjusting for grade inflation and/or grading inconsistency. Accordingly, this section of the paper will demonstrate the efficacy of one of those techniques—linear adjustment. This efficacy will be shown through a

simulation where grades for entire departments are raised or lowered in order to create substantial inconsistency[9]. The question then is to see if linear adjustment is accurately able to sort out this artificially imposed "noise". Data for this section is drawn from the Macalester College Registrar. The data set includes 442 students who graduated as part of the Class of 2005 at Macalester.

Before delving into the analysis of linear adjustment, a brief review of the method is in order. The method works by calculating a predicted grade according to a linear fit of a latent ability parameter.

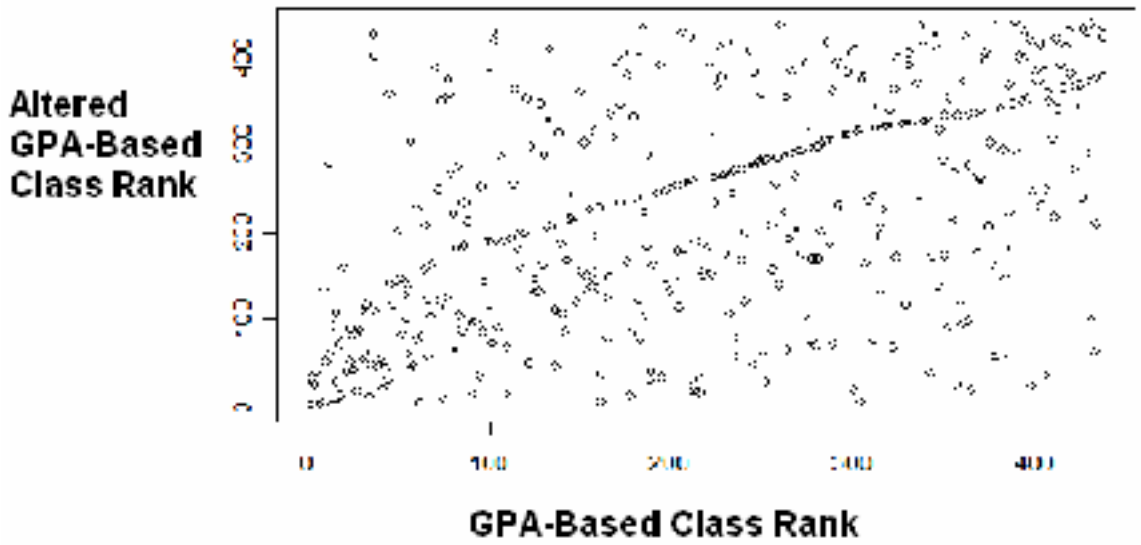
$$\text{Predicted Grade} = \alpha * \theta + \beta$$

The method then seeks to minimize the squared error between this predicted grade and the actual grade received by each of the students within the course. Because both student ability (θ) and the course parameters (α and β) are unknown, this fit must be done iteratively. The initial value used within this iterative fit is that we assume that student ability is equal to GPA-based class rank. This, of course, is not a very accurate measurement of student ability but error in this initial estimation will rapidly be eliminated by the iterative process.

In the simulation, all grades within the Chemistry department were raised by three points, all grades in the Psychology department were raised by two points, and all grades in the Math and Biology departments were lowered by three points. This alteration quite effectively skewed GPA-based student rank such that there appears to be little correlation between a student's new GPA-based class rank and the original GPA-based class rank. In the following scatterplots, student performance is measured by student ranking where 442 represents the highest student ranking and 1 represents the lowest student ranking.

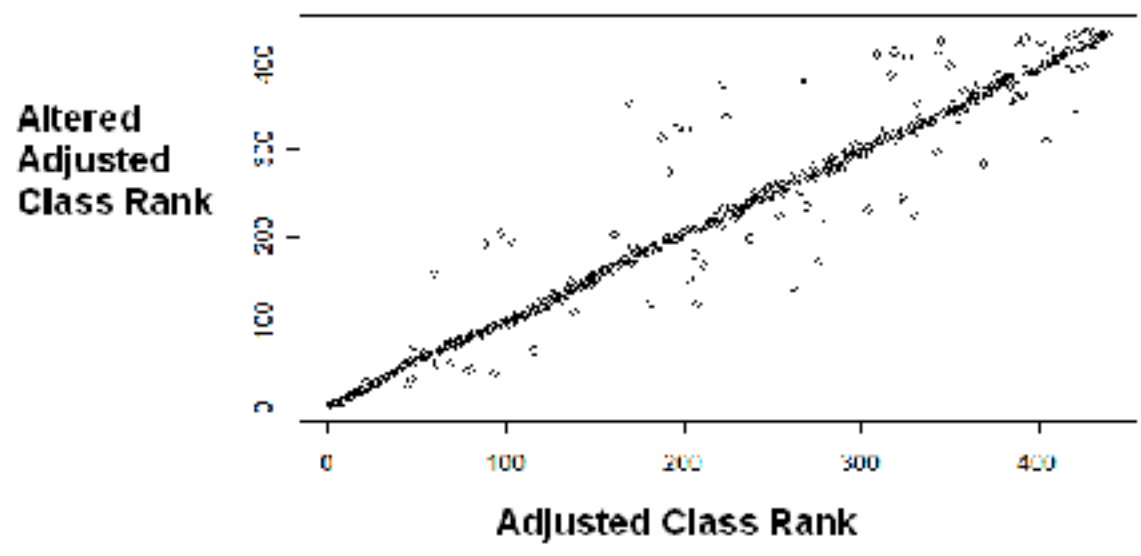
[9] This simulation was done jointly with Professor Danny Kaplan at Macalester College. Programming for the simulation was done using the statistics package, R.

Figure 15: Scatterplot showing correlation between GPA-based class rank prior to and following experimental imposition of the inconsistency.



However, application of the linear adjustment method to this grade data greatly decreases the spread imposed by the addition of this grading inconsistency. The method appears to have moved points which were skewed substantially one way or another back toward the line where the unaltered rank is equal to the altered rank (along this line there is no inconsistency).

Figure 16: Scatterplot showing correlation between adjusted class rank prior to and following experimental imposition of the inconsistency.



This correction is quite impressive given the magnitude of the inconsistency imposed. When one considers that inconsistency in actual grade data is rarely so extreme, linear adjustment method is only likely to perform better when applied to such data.

The simulation provides strong support for the use of linear adjustment of grades in determining class rank. It shows that the method is able to correct for a substantial portion of grading inconsistency, providing a more accurate understanding of student performance. Thus, the question is not whether it is possible to correct for grading inconsistency, but simply what the best method is to do so.

VII. Conclusion

A. Summary

This paper identifies three primary reasons why adjustments for grading inconsistency are necessary. First, unadjusted grades provide a misleading evaluation of student performance which can produce artificial pressure on student course selection. Second, unadjusted grades make outside evaluation of academic performance difficult. Potential employers are deprived of the information they need to choose the most qualified candidate, and are left to rely on 'old boys' networks to select employees, further disadvantaging traditionally disadvantaged groups. Third, a lack of grade adjustment increases pressure on professors to assign high grades in exchange for positive course evaluations and higher course enrollments in future semesters. This process can undermine professorial autonomy in assigning grades, decreasing the ability of grades to serve as an effective motivational tool.

In response to these shortcomings, a handful of universities around the country have taken action to address grade inflation and/or grading inconsistency. These efforts have had

mixed success; most see little or no change in grading practices post-reform. The problem is that most reforms require changes in the practice of assigning grades itself. This is a weaker approach because it both fails to account for the pressures which professors face in assigning grades and can limit the freedom of professors to assign grades in order to optimize their classroom environment.

More mathematical approaches to grading inconsistency, such as calculation of "Real GPA", the "Achievement Index", and linear adjustments, have also been proposed. Real GPA and linear adjustments suffer from limitations which ensure that they will inaccurately correct for grading inconsistency, either providing a misleading evaluation of student performance or creating disincentives to take particular courses based upon the grade distribution given by that particular professor. On the other extreme, the Achievement Index has been rejected for its mathematical complexity, which engenders opposition when implementation at a college or university is proposed.

Within this context item response theory (IRT) is offered as a solution to grading inconsistency. IRT would not require any change in grading practices; it would simply provide an adjusted ranking to aid in interpreting grades which have already been assigned. Moreover, IRT is not subject to the criticisms of bias in adjustment leveled against Real GPA and linear adjustments. Finally, IRT is a more acceptable mathematical technique. It is based in broadly accepted test theory, decreasing the odds that it will meet opposition in implementation due to mathematical complexity.

B. Recommendations

Based upon the preceding analysis, I believe that Macalester would benefit from using IRT to adjust for grading inconsistency. The change would be relatively minor:

transcripts would simply carry an additional piece of information, the student's IRT-based adjusted class rank. The rest of the transcript would remain unchanged. While the adjustment process would be simple, the implications would be notable. With academic performance better contextualized, the disincentive to take more severely graded courses would be reduced or removed altogether. This adjusted class rank would more accurately reflect student performance, better informing potential employers. Professors would likely receive less pressure from students to distribute high grades as relative standing (to other students) in a course would be more important than absolute standing (the grade itself).

Beyond these improvements in the educational process, Macalester would stand to benefit from taking a leading role among its cohorts in adjusting for grading inconsistency. The schools which have taken action to address, primarily, grade inflation (such as the University of Colorado, Harvard, and Princeton) have received extensive, mostly positive, publicity. The public perception is that addressing problems with grading is a forward-thinking strategy which shows that the school is both daring and shrewd. Macalester, however, has the opportunity to advance beyond the basic approaches tried at other schools and implement a policy which focuses on grading inconsistency, instead of just grade inflation and avoids the pitfalls of less appropriate grading remedies.

Bibliography

Arenson, Karen W. (2004), "Is it Grade Inflation, or Are Students Just Smarter?" *New York Times,* April 18.

Baker, Frank B. (1992), *Item Response Theory: Parameter Estimation Techniques*. New York: Marcel Dekker.

Baker, Frank B. (2001), *The Basics of Item Response Theory* (2nd Edition). Published online by the ERIC Clearinghouse on Assessment and Evaluation. Available at: http://edres.org/irt/baker/final.pdf.

Berry, Scott M. (2006), "Does Defense Win Championships." *Chance*, Vol. 19, No. 2, p. 52-54.

Berube, Michael. (2004), "How to End Grade Inflation." *New York Times*, May 2.

Boston Globe. (2004), "Harvard Still Home of Easy A." February 15.

Brown, Jennifer. (2006a), "Brown assails grade inflation: CU president wants to put class rank or grade percentiles on student transcripts The proposal may give potential employers a clearer picture of a student's achievement. But some faculty take issue with the plan." *Denver Post*, August 31.

Brown, Jennifer. (2006b), "Inflation in grades debated by CU regents." *Denver Post*, September 7.

Brown, Jennifer. (2006c), "Regents opt to reveal rankings The CU board adopts the smallest measure offered to remedy grade inflation: class rank, by request only." *Denver Post*, December 6.

Caulkin, Jonathan P., Larkey, Patrick D., and Wei, Jifa. (1996), "Adjusting GPA to Reflect Course Difficulty." Working Paper, Heinz School of Public Policy and Management, Carnegie Mellon Univ. Available online at: http://www.heinz.cmu.edu/wpapers/detail.jsp?id=35

Clayton, Mark. (2002), "Amid Cries of Grade Inflation, C's Still Abound." Christian Science Monitor, August 20.

Crenshaw, Albert B. (2002), "At Harvard, a Steeper Grade." *Washington Post*, May 26.

Elliot, Rogers and Strenta, A. Christopher. (1988), "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly." *Journal of Educational Measurement*, Vol. 25, No. 4 (Winter), p. 333-347.

Goldman, Roy D., Schmidt, Donald E., Hewitt, Barbara Newlin, Fisher, Ronald. (1974), "Grading Practices in Different Major Fields." *American Educational Research Journal*, Vol. 11, No. 4 (Autumn), p. 343-357.

Golman, Roy D. and Widawski, Mel H. (1976), "A Within-Subjects Technique for Comparing College Grading Standards: Implications in the Validity of the Evaluation of College Achivement." *Educational and Psychological Measurement*, Vol. 36, p. 381-390.

Healy, Patrick. (2001), "Harvard's Quiet Secret: Rampant Grade Inflation". *Boston Globe*, October 7.

Healy, Patrick. (2002), "Harvard to Award More B's, Raise Honors Standards." *Boston Globe*, May 22.

Jeffrey, William. (2006), "Importance of Basic Research to United States' Competitiveness." Testimony before the Committee on Commerce, Science, and Transportation. Subcommittee on Technology, Innovation and Competitiveness. United States Senate. March 29.

Johnson, Valen E. (1997), "An Alternative to Traditional GPA for Evaluating Student Performance." *Statistical Science*, Vol. 12, No. 4 (Nov.), p. 251-269.

Johnson, Valen. (2002), "An A Is an A Is an A…and that's the problem." *New York Times*, April 14.

Johnson, Valen. (2003), *Grade Inflation: A Crisis in College Education*. New York: Springer-Verlag.

Kohn, Alfie. (2002), "The Dangerous Myth of Grade Inflation." *The Chronicle of Higher Education*, Vol. 49, Issue 11 (November 8).

Kurtz, Stanley. (2006), "Deflating Grade Inflation." *National Review*, September 27.

Leonard, David K. and Jiming Jiang. (1999), "Gender Bias and the College Prediction of the SATs: A Cry of Despair." *Research in Higher Education*, Vol. 40, No. 4, p. 375-407.

Lord, F.M., and Novick, M.R. (1968), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mansfield, Harvey, C. (2001), "Grade Inflation: It's Time to Face the Facts." *The Chronicle Review*. http://chronicle.com/free/v47/i30/30b02401.htm

Merrow, John. (2006), "Grade Inflation: It's Not Just an Issue for the Ivy League." *Carnegie Perspectives*. http://www.carnegiefoundation.org/perspectives/sub.asp?key=245&subkey=576&printable=true

Meyer, Meredith. (2005), "Students get a .76 boost from grade inflation." *Chicago News*, January 18.

Mount, Harry. (2005), "Ivy League university cuts 'A' quota to halt grade inflation." *The Daily Telegraph (London)*, September 22.

Neal, Anne D. (2006), "CU gets an 'A' for grade inflation fight." *Rocky Mountain News*, September 25.

Partchev, Ivailo. (2004), "A visual guide to item response theory." Available online at: http://www.metheval.uni-jena.de/irt/VisualIRT.pdf.

Rosovsky, Henry and Hartley, Matthew. (2002), "Evaluation and The Academy: Are WeDoing the Right Thing?" American Academy of Arts and Sciences.

Shephard, Alicia. (2005), "A's for everyone; in an era of rampant grade inflation, some college students find it shocking to discover that there are 26 letters in the alphabet." *Washington Post*, June 5.

Strenta, A. Christopher and Elliot, Rogers. (1987), "Differential Grading Standards Revisited." *Journal of Educational Measurement*, Vol. 24, No. 4 (Winter), p. 281-291.

USA Today. (2002), "Ivy League Grade Inflation." February 8.

Washington Post. (2004), "At Princeton, a Move Toward Fewer A's; Professors Vote for New Policy Aimed at Reversing Trend of Grade Inflation." April 28.

Young, Jeffrey R. (2003), "Researchers Charge Racial Bias on the SAT." *Chronicle of Higher Education*, Vol. 50, No. 7, p. A34-A35.

Young, John W. (1990), "Adjusting Cumulative GPA Using Item Response Theory." *Journal of Educational Measurement*, Vol. 27, No. 2 (Summer), p. 175-186.

Young, John W. (1993), "Grade Adjustment Methods." *Review of Educational Research*, Vol. 63, No. 2 (Summer), p. 151-165.