Management by Results: Student Evaluation of Faculty Teaching and the Mismeasurement of Performance

Laura Langbein
School of Public Affairs
American University

For presentation at Annual Meeting of Public Choice Society, Mar. 10-13, 2005, New Orleans, La.

Abstract

Management by Results: Student Evaluation of Faculty Teaching and the Mismeasurement of Performance.

This study uses data on all courses at American University over a four-year period to test the hypothesis that faculty who give higher grade get better SETs. Regression results show that actual grades have a significant, positive effect on SETs, controlling for expected grade and fixed effects for both faculty and courses. The results do not change if possible endogenity between SET and actual grade is accounted for.

One implication is that students, faculty, and provost are engaged in an individually rational but arguably socially destructive game centered on the link between SETs and grades. The overall social effect is to make both the SET a faulty signal of teaching quality and grades a faulty signal of future performance on the job. A more general implication for public management is that, when performance is hard to measure, using piece-rate pay-for-performance systems, embodied by the linear relation between SETs and faculty pay, may have unintended adverse consequences. Performance is typically hard to measure in many non-profit and public organizations. It may be advisable to design a more open contract based on discrete pay systems and self-enforced norms to assure good performance in organizations such as these.

Introduction and Background:

The asymmetric relation between principals and agents has long been used to characterize the relation between managers and employees (Holmstrom, 1979; Baker, 1992). In this relation, employees work for (higher) pay, and managers want to induce desired behavior by the employees. But the employees know more about their behavior than the manager, who cannot directly observe everything the agent does. The consequence is that the employee is likely to shirk, either by doing nothing at all, or by engaging in "sabotage," which is work that the principal does not favor (Brehm and Gates, 1997). The dilemma is exacerbated when the output desired by the manager must be jointly produced by multiple agents, because it is not possible to design an efficient pay system under these circumstances (Holmstrom, 1979; Baker, 1992; Holmstrom, 1982; Gibbons, 1998; Kerr, 1975). The imbalance is magnified in horizontal, not-for-profit organizations such as universities, where centrally located managers are unable to observe or even measure the output of the core, specialized, street-level employees, namely the university faculty. University administrators are under increasing pressure by Boards of Trustees and political overseers to show increasing productivity. Higher education costs appear to be rising faster than measurable output (e.g., number of degrees granted) (Levin, 1991). While universities may be victims of Baumol's "cost disease," that excuse will not satisfy taxpayers or those who contribute to the university voluntarily. While it appears to cost more and more to put a faculty member in front of a classroom, one simple way for administrators to control these costs is to more closely link faculty pay with faculty performance. While administrators cannot directly increase faculty productivity with innovative new technology, they can increase productivity by reducing or preventing increases in costs (i.e., faculty salaries) when there is no increase in measurable output.

The two main activities of university faculty are teaching and research. In U.S. universities (and many others, including Canadian universities), evaluation of research output (that is, whether and where it is published) remains the province of review by peers outside the university in which the faculty member is employed, and thus largely outside the purview of university administrators. But teaching remains internally controlled by university managers and evaluated within the university. However, compared to "good" research, "good" teaching is particularly hard to observe and measure.

In this context, traditional principal-agent theory assumes that complete contracts are necessary to prevent employees from shirking (Baker, 1992). Complete contracts reward specific behaviors with higher pay. Direct observation of classroom teaching by supervisors (or faculty peers) would be one way to enforce such a contract, but this requires costly monitoring. By contrast, direct observation of faculty research by peers is less costly, because the marginal cost of peer review enhances the quality of the refereed papers. It may also enhance the quality of the reviewer's own research and reputation, reducing the net costs of the monitoring process.[1] In some cases of research, monitoring peer output may actually be a net benefit rather than a net cost. In most cases, monitoring employee output is regarded as a net cost. While monitoring research by direct

[1] For a discussion of the economic efficiency of the process of peer-reviewed research, see Thornton (2004).

observation may be an exception to this rule, monitoring teaching by direct observation is not an exception. Managers prefer a less expensive form of monitoring what goes on in a classroom (or in the faculty office or in the laboratory). Student opinions of faculty teaching have become the preferred, low-cost mechanism by which university administrators monitor faculty teaching, prevent "shirking," and maintain productivity (Seldin, 1993a; Kulik, 2001; Haskell, 1997; Ory, 2001; Ory and Ryan, 2001; Gray and Bergman, 2003; Becker, 2000).[2]

More recently, theorists (Gibbons, 1998; Chen, 2002; Hartwig, 2004; McKeachie, nd; Frey and Oberholzer-Gee, 1997; Kreps, 1997; Frey, 1993; Nagin et al. 2002) have also suggested that, when output is hard to measure and the production function is unknown, incomplete (unspecified) contracts between employees and managers may be more efficient than complete contracts.[3] In the context of university teaching, the Student Evaluation of Teaching (SET) embodies a complete "piece-rate" contract for output. That is, in most colleges and universities, aside from research, faculty are rewarded with promotion, tenure, and higher pay for "better" teaching, as measured entirely or in large part by the SETs (Kuklik, 2001; Seldin, 1999; Haskell, 1997; Grimes et al., 2004). The SET is assumed to be a usable, generally accurate index of teaching "quality."

If this assumption is faulty, then the widespread use of SETs to ascertain individual performance may be neutral at best, if it is relatively inexpensive and only drives out the use of more costly "better" measures. If the use of SETs as a low cost measure of performance not only drives out better measures but also causes unwanted behavior, then the use of SETs is not just wasteful but even more costly as well (Baker, 1992). In this vein, it is commonly alleged that SETs are one (and perhaps the main) cause of student grade inflation.

---

[2] Many studies lead directly or indirectly to the conclusion that student evaluations of faculty teaching are consequential, not only for individual faculty, but for the joint or social product of higher education more generally, including academic freedom (Haskell, 1997). Hamermesh and Parks (2005) provide a recent example. They argue that student evaluations are used for faculty pay determination and promotion, regardless of whether the evaluations correspond to any measure of underlying quality of teaching. Sproule (2002) argues that when student evaluations are used as part of faculty pay determination, they may need to be adjusted for factors not under the control of faculty. Sproule (2000) points also to the public good aspects of higher education, and notes that students evaluate teaching as consumers, ignoring the social externalities of higher education. He, as well as Gray and Bergman (2003), argue further that student evaluations level the difference between students and their faculty-teachers, degrading the authority of teachers to teach, and teaches students to value mediocrity and immediate clarity. Sproule also documents the common use of student evaluations in Canadian universities.

[3] Most theory (and what limited empirical evidence there is) supports the expectation that, in norm-driven organizations, extrinsic rewards crowd out intrinsic rewards. For example, Hartwig (2004) notes the mismatch between "utilitarian" rewards (e.g., merit pay) in a normative organization. He and others (e.g., Frey, 1993; Kreps, 1997) point to the importance of implicit control mechanisms in universities rather than explicit piece-rate pay-for-performance schemes. McKeachie (nd) speculates that using student evaluations as a continuous, explicit measure of teaching merit is demoralizing, and may make teaching worse. Using an experimental design, Nagin et al. (2002) find that employees who "trust" their employer and view the employer positively cheat less and are less responsive than others to increases in monitoring (i.e., direct controls). The implication is that it is costly for employers to monitor "good," "trusting" employees, because it adds more to their costs than to productivity benefits. In the case of a university, basing faculty pay on student evaluations may do the same by adding to costs and possibly decreasing teaching productivity. Even in the case of primary schooling, where student achievement is commonly measured by standardized tests, merit pay, at least in the Tennessee STAR study, had mixed success in rewarding the teachers who increased student achievement (Dee and Keys, 2004).

The research presented below uses data on all courses at American University over a four-year period to test the hypothesis that faculty who give higher grades get better SETs. There are many other determinants of both grades and SETs, and these factors must be identified and held constant to isolate and validly estimate the impact, if any, of grades on SETs. These factors, identified in previous research, include teacher characteristics unrelated to teaching ability (e.g., age, sex, physical attractiveness, etc.); student characteristics (level of motivation; prior ability; prior education in the field of the course, etc.); and course characteristics (size; whether it is required or not; when the course is offered; etc.).[4] Rather than try to identify and measure all of these variables individually, an impossible task in a single study, this research uses both faculty and course fixed effects as control variables that act as summary surrogates for the entire set of course and faculty characteristics that potentially affect SETs. Since the data are course level data, the course fixed effects control also for student characteristics as they vary among classes. The only measured variables reflect grades and SETs: the actual average grade in the class, the average of the expected grades, and the average SET in the class (itself an average of the course and instructor rating, which are highly correlated).

The basic estimating equation is:

$$SET_i = a + b\ AverageCourseGrade_i + c\ ExpectedGrade_i + \Sigma_k d_k\ D1_{ik} + \Sigma_l d_l\ D2_{il} + e_i,$$

where i = 1,…,N classes, over 8 semesters
k = 1,…,K unique classes
l = 1,…,L faculty members
$K*L \leq N$

In this formulation, the null hypothesis is that the parameters estimated by coefficients b and c are 0; the alternative is that the parameters are positive, when faculty and course fixed effects are included.

The study also investigates the exogeneity of actual average course grades with respect to the course SET. The results show that actual and expected grades both have a significant, positive effect on SETs, controlling for faculty *or* course fixed effects, or for faculty *and* course fixed effects. The results also suggest that, while average course grades may indeed be endogenous, the endogenity-adjusted estimates consistently indicate an even larger effect of actual grades on SETs. Expected grades may even have

[4] There is a voluminous literature on the determinants of SETs. For example, d'Apolonia and Abrami (1997) argue that SETs are influenced by administrative and instructor characteristics, and by the course grade. They view the SETs as moderately valid indicators of teaching quality. Greenwald and Gilmore (1997) point to the importance of workload as a determinant. Langbein (1994) also finds that numerous faculty, class and student characteristics affect SETs, as does McPherson (2005) and Boex (2000), Hoyt and Pallett (1999), and Hamermesh and Parker (2005). Boex emphasizes the importance of student perceptions of clarity and organization (not the same as teaching so that students learn), while Hamermesh and Parker find that looks matter. Grimes et al. (2004) find that psychological predisposition is important, among other variables: students who feel controlled by external forces are more likely to blame the professor. Hoyt and Pallett (1999) summarize much of the earlier literature. Stapleton and Murkison (2001) examine whether SETs reflect (are determined by) what they are supposed to measure (studying and learning), or if they are affected by expected grades. I consider this in more detail below.

no direct effect. The endogeneity tests do not yield clear evidence concerning the issue of whether high SETs produce more or less learning.

The overall implication is that students, administrators and faculty are engaged in an individually rational but arguably socially destructive game. Administrators want higher SETs because it leads to higher grades and higher student retention rates, which means more tuition and tax revenues. Faculty want higher teaching evaluations because it leads to higher salaries, and students want higher grades for the same reason. But the overall social effect is to make both the SET a faulty signal of teaching quality and grades a faulty signal of future performance on the job. No student, no individual faculty member, no individual college or university administrator, and no college or university acting collectively has much of an incentive to break this vicious cycle.

Pay for performance systems are based on the theory that, if you pay for A, you will get A. But, when performance is hard to measure, A may not be what you really want, and may have socially destructive outcomes as well. This may be particularly true in rank order markets, where relative position is what counts. Rank order markets induce excessive effort to improve whatever is ranked. In the case at hand, grades are ranks, and SETs are too, but they need not be. When SETS are used to rank faculty as part of merit pay determination or other personnel decisions, they create rank order markets within universities. Rethinking how SETs are used may be a vehicle to break the vicious cycle.

I model the three players in higher education with simple net payoff equations.[5] The president or provost is the principal in this organization (or represents the trustees or legislators, who are the ultimate principals). The faculty are the agents, who respond to their own preferences, and to those of the principals. Finally, the students are the customers. In my formulation, each of the equations is linear. A more realistic model would represent what each player values with benefits that diminish at the margin; it would represent what each player does not value as costs that increase at the margin. But this modification would only add complexity with no additional information.

Theory: The agents (faculty)

Faculty are rewarded in a complete contact for certain specific behaviors (Lazear, 2000).[6] (I ignore the role of service.) Specifically, they are rewarded for higher SETs.[7] Higher SETs are not necessarily the same as "better teaching." (I address that issue below.) Faculty are also rewarded for scholarly output. In contrast with teaching, I assume that scholarly output is measured without significant random error, and with no systematic error. Pay may be continuous or discrete (Lazear, 2000). Merit pay is usually continuous pay, usually based on ordinal rankings, but it may also be discrete. In

[5] Stone (1995) suggests such a setup, with no formal model. Kanagaretam et al. (2003) develop a formal model. The logical result shows that overemphasis on SETs, especially when student quality is not at the highest level, results in higher grades and less effort by students directed at learning. Their model, however, omits the provost (or any principal) as an actor.

[6] Boylan (2004) shows, in a completely different context, that agents (federal prosecutors) respond to performance inducements even when the link between the output indicator (longer prison sentences) and pay is loose and implicit

[7] Haskell (1997) notes that over 80% of faculty report this in formal surveys. See also Seldin (1993a, 1999).

contrast, promotion and tenure reflect discrete pay categories, where you are either over the bar, or not.

The faculty pay equation is:

$$\text{Net payment in year } t = P_t = \alpha_1 q_1 + \alpha'_1 q_1 (q_1 > q_{1\,min}) + \alpha_2 q_2 + \alpha'_2 q_2 (q_2 > q_{2\,min}) + \gamma P_{t-1} - c_1 q_1 - c_2 q_2$$

where q_1 = measured SET ($\neq q'_1$ = "real" teaching quality)

q_2 = measured research productivity (= q'_2 = "real" research productivity)

$\alpha_{.}$ = continuous reward response (merit pay) for SETs, research

$\alpha'_{.}$ = discrete reward response (promotion, tenure pay) for SETs, research

$c_{.}$ = cost of effort (opportunity cost and disutility) for SET-based teaching, research (one cost may be higher than the other).

γ = weight of previous (base) pay.

$\gamma > 1.0$ => across the board raise, with additional (or no) rewards based on research/teaching

$\gamma = 1.0$ => no across the board raise; all raises based on research/teaching

$\gamma < 1.0$ => base pay can be cut

My model assumes that rewards are NOT based on effort. Effort is omitted entirely, because, for faculty, effort is assumed to be unobservable. Faculty pay systems, in my model, are based on output alone (Lazear, 2000). I also assume that there is no error in peer review of faculty research ($q_2 = q'_2$).[8] In a complete contract, all raises are based as a continuous function of increased research and teaching. In an entirely open contract, raises are across the board: $\alpha_{.} = \alpha'_{.} = 0$.[9] Note that measurement accuracy (or inaccuracy) of either teaching or research is irrelevant when contracts are open. Note also that the monetary payment is not the net pay. I deduct from the monetary pay the disutility that individual faculty may have for teaching (especially when it is directed at raising SETs) and research. Thus, the value that faculty might attach to "quality" teaching is included as part of the disutility (c_1) that faculty feel from having to produce high SETs

[8] England (nd) makes a similar argument. See also Thornton (2004).

[9] Hartwig (2004) cites traditional public administration literature that distinguishes between theory X and theory Y; in norm based, theory Y, organizations, raises "should" be across the board.

in exchange for "quality" teaching. (This assumes that c_1 is positive; it can also be negative, if SETs and "quality" teaching are complements.) The actual monetary pay is ($P_t + c_1 q_1 + c_2 q_2$). This term also appears in the principals payoff equation.

Theory: the principal (the president or provost)

The provost's immediate payoff is from net university revenues (Stone, 1995). Revenues come from tuition (paid either directly by students or taxpayers), from external research grants or contracts, and from alumni donations. The largest deduction from revenues is for faculty salaries. Tuition revenues are easiest to generate when students do not transfer; as a consequence, provosts value student retention, because more revenues are retained when students are satisfied and choose not to leave the institution early, and when students are not dismissed because of low grades. Students indicate their probable behavior with SETs (q_1). High SETs signify to the provost that students are satisfied, and are not likely to leave.

It is particularly costly if students leave early. If these students leave, they must be replaced with new students. It is more costly to unnecessarily recruit a new student to replace, say, a freshman or sophomore, than it is to retain the old student. Thus, revenues from old, or continuing, students (R_{old}) have no deduction for recruiting costs associated with revenues from new students ($R_{new} - C_{new}$). Retaining current students is directly facilitated by high SETs, and by not giving low grades.[10] High SETs are (hypothetically) facilitated by higher grades, and by avoiding low grades. Thus, grades have a direct impact on retention revenues when poor students are not flunked out, and an indirect impact by increasing student course satisfaction (measured by SETs). Revenues also come from new students, who must be recruited. Recruiting students to the base is a fixed cost, but recruiting new students to replace current students who leave voluntarily or involuntarily is an unwanted variable cost. Another source of revenues is faculty research ($R_{res} = R_{res}(q_2)$), because it generates grants and contracts; faculty research may also contribute to university prestige and consequent alumni donations. Of course, there are other miscellaneous sources of revenues, including sports and arts programs, university enterprises (e.g., the campus store, the university press), and the like. The full equation is:

$$V_t = a_1 R_{old} + a_2 (R_{new} - C_{new}) + a_3 R_{res} + a_4 R_{misc} - (P_t + c_1 q_1 + c_2 q_2) \geq V_{t-1}$$

$$R_{old} = f(R_{old}, R_{old}(S_t))$$

$$R_{res} = R_{res}(q_2)$$

In other words, to generate more revenues than the previous year's base (V_{t-1}), the provost must first deduct from revenues the monetary payments to faculty ($P_t + c_1 q_1 + c_2 q_2$), which I assume is a major cost to the university. While those payments for faculty salaries rise with merit and promotion pay for more research and higher SETs, university revenues also rise with more faculty research and higher SETs. So the pay system is roughly incentive compatible, at least from the perspective that both faculty and

[10] Haskell (1997) also mentions the importance of SETs for retention.

administrators benefit from research and SETs. But the value that faculty and administrators attach to these activities is unlikely to be the same. Higher SETs enter into the provost's payoff indirectly through the revenues that come from retaining current students, or R_{old} , which is a direct function of S_t, the students' value or satisfaction function as defined in the next section. The provost wants to see high SETs, because retention increases as SETs increase, at least up to a point (Langbein and Snider, 1999; Haskell, 1997; Crumbley and Fliedner, 2002). The provost also values high quality scholarly productivity, because research, based on competitive external peer review, does generate more external grants and possibly contracts too, as well as alumni donations for named chairs and research institutes, and for the recognition of association with a prestigious university or college. The equation also makes clear that the provost, along with the students, whom I discuss below, values high grades. This is an induced preference; that is, the provost wants to retain current students, and retaining current students is promoted by providing what students value, measured by S_t. As we see below, S_t depends directly, in part, on both grades and SETs. Hence, to enhance retention, the provost wants what the current students want.

Theory: the consumer (the student)

Students value three things from a college education. First, they value grades, because good grades lead to more and better post-graduate job and educational opportunities. Second, students like to enjoy the courses they spend time and money to take; they want to be satisfied consumers. Third, some (perhaps most) students value the process of learning, because of its intrinsic value, or because it has long-term extrinsic value in the market.[11] Finally, there is an opportunity cost of effort, represented not only by tuition and foregone wages but also the disutility of studying. The full equation is:

$$S_t = \gamma_1 g + \gamma_2 q_1 + \gamma_3 q'_1 - c E.$$

In this model, S_t is the net value of a year of college education to the student. The student's GPA is g; the responsiveness of a student's value to grades is γ_1. The coefficient γ_1 could be 0 or negative, but it will usually be positive. The measured SET is q_1, and the responsiveness of a student's value to the measured SET is γ_2. Some students value their satisfaction with a course more than others. In fact, some students may even regard "fun" courses negatively (Kanagaretnam, 2003), so it is possible that $\gamma_2 < 0$. Finally, some students value the "true" (unmeasured) instructional quality of a course (q'_1); for these students, $\gamma_3 > 0$. For students who place no value on the underlying instructional quality of a course, $\gamma_3 = 0$. It is possible, but not likely, that quality gives some students disutility ($\gamma_3 < 0$). Finally, E is a summary of the costs of education; these

[11] Pollio and Beck (2000) provide evidence that students value both grades and "learning." The learning component may be related to the public good aspect of higher education, or it may represent the intrinsic value of learning for the student, and that may not have a spillover. In other words, q' can be either a private good (value of learning to the student) or it could be the private value for a good that also has public value. See Sproule (2000).

include both monetary and non-monetary costs, and the coefficient reflects the students' responsiveness of demand with respect to increases in these costs.

Three things are notable about this set of three equations. First, only the students (or some of them) value (unmeasured) instructional quality. Neither the provost-principal nor the faculty-agent receives any immediate, direct payoff from good quality teaching, while some students clearly do.[12] Second, both the students and the provost value grades. The students preference for high grades is quite direct; higher grades are likely to lead to better future outcomes. The provost's preference is indirect. The provost values higher grades because higher grades aid in retaining current students. Higher grades prevent students from leaving involuntarily (from low grades) or voluntarily (because high grades contribute to higher SETs, and overall student satisfaction). (I show below that faculty may indirectly value grades; this would be the case only if grades determine SETs.) Third, the only variable that is clearly common to each of the three payoff functions is SETs, or "measured" instructional "quality". That is, the measure of instructional quality (q_1, the SET) is the only variable that appears in all three equations. In effect, it is the currency that binds students, administrators and faculty. Just as prices of products bind consumers, firm employees, and firm managers to one another, SETs are, in effect, a common currency of university life, and function, perhaps even more than grades, as price signals in the university economy. Figure 1 summarizes these overlapping preferences. But it raises a question about the centrality of grades in the university economy.

Figure 1: Three Actors and Three Values

The Values

	Grades		SETs (Measured "Quality" Instruction)		Instruction	
The Actors						
Students	√		√		√	
Faculty		√ ??		√		√
Provost	√		√			

Critical to the equations behind the three value functions is the relation between grades, q_1 (the SETs) and q'_1("quality" instruction). That is, are SETs a reliable and valid measure of instructional quality? If they are, then we can ignore their inequality in the student equation and the absence of q'_1 in the provost equation. If SETs were a valid representation of instructional quality, then all of the actors would value it. Further, grades would only be valued by the students and the provost. That is, grades would not drive a valid measure of instructional quality (though the reverse could well be true, since good instruction facilitates learning). The provost would still prefer to avoid poor grades in order to retain students, but would be incapable of punishing faculty if student

[12] Faculty indirectly value "good" teaching by reducing the disutility with which they regard teaching directed at maintaining or improving SETs.

performance really warranted a low grade. Students would also prefer to avoid low grades, and prefer to receive higher ones, but they will no longer be able to punish faculty for giving low grades. Thus, grades would become a more accurate signal of student performance in the external market and in the internal market for course selection.

If, on the other hand, SETs are partly determined by grades, then grades would become (indirectly) valued by faculty. Faculty will be motivated to give higher grades to get higher SETs. This is particularly likely because an instructor's skill set is not very manipulable. A poor instructor cannot easily just "teach better" to make her evaluations go up. Thus, given relatively fixed inputs at the individual level of teaching, something else needs to be manipulated that is a) easier to change and b) translates readily into a recognizable medium of exchanges between the instructor, the students, and the provost. Grades provide that medium of exchange.[13] If grades affect SETs, then all three actors will value both high SETs and high grades, and they will in effect all be exchanging one currency for the other. Faculty give students high grades to get high SETs. Students reward the provost with high retention in return for high SETs and grades. And the provost rewards the faculty with high pay in return for high SETs (and consequent higher grades). Thus, the relation between grades, "quality" and "measured" instruction is critical to understanding the political economy of the university.

Instructional Quality, and SETs and Grades

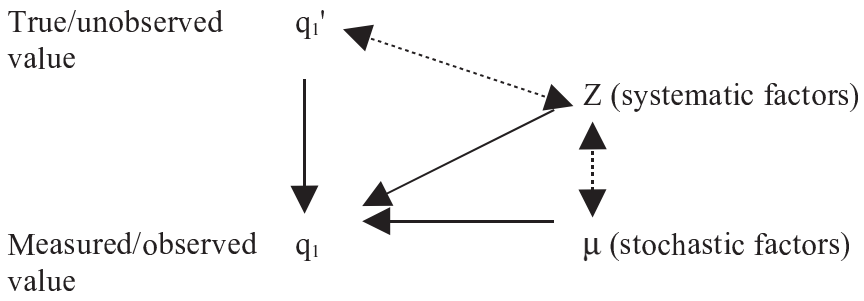
A considerable amount has been written on this issue, and I do not review the specific details here, but I do summarize what appears to be an emerging consensus. First, I note two traditions of measurement, one from psychology and education, the other from other social sciences. Both focus on reliability and validity, but the former refers to reliability as "convergent" validity and validity as "divergent" validity. A more precise definition is that a reliable measure has a relative absence of random measurement error, while a valid measure has a relative absence of non-random measurement error. (Sproule, 2002 also uses these criteria in his review of the literature on the reliability and validity of SETs.)

Figure 2 below is a conceptual diagram of these twin concepts, in the context of the relation between q'_1, the "true" value of the concept of instructional quality, and q_1, the measured or observed value of instructional quality. The presence of random measurement error (or unreliability) would be signified by the magnitude of the impact of μ on q_1; if its impact is relatively "low," then q_1 (measured performance) would be a reliable indicator of q_1' (unmeasured or "true" performance) (Langbein, 1994). The presence of non-random measurement error (or invalidity) would be signified by the relatively large impact of Z, a vector of outside variables (such as class size, faculty age or gender, or student grade) conceptually unrelated to true faculty performance (q_1'), on measured performance (q_1). Non-random error might also be signified by a consistent over- or under- estimation of the measured score, indicated by a non-zero intercept. Alternatively, an invalid indicator would be signified by a low correlation between measured and unmeasured performance, commonly estimated by the use of principal

[13] I thank David Levin for this point.

components analysis or confirmatory factor analysis when there are multiple measures of performance.

Figure 2: The relation between "true" and measured instructional quality.



The equation for the diagram in Figure 2 is:

$$q_1 = \beta_0 + \beta_1 q'_1 + \beta_2 Z + \mu$$

This equation helps to clarify the difference between reliability and the two types of validity. Reliability refers to the expected variance of μ; convergent validity refers to the expected value of β_1; and divergent validity refers to the expected value of β_2 and β_0. Table 1 below summarizes the expected values of these parameters.

Table 1: The statistical implications of a reliable and valid measure:

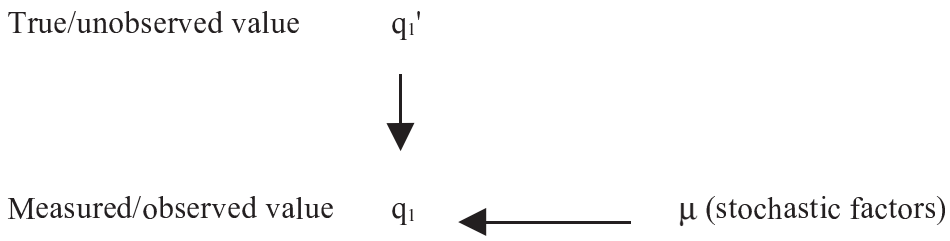
Reliability (absence of random measurement error)	Validity: Convergent (does it measure what it is supposed to)	Validity: Divergent (absence of non-random error)
$E(\mu_i^2) \rightarrow 0$	$E(\beta_1) \rightarrow 1$ or $E(\beta_1) > 0$	$E(\beta_2) \rightarrow 0$ $E(\beta_0) = 0$

The issues of reliability and convergent validity concern the relative importance of the stochastic term, compared to the unmeasured "true concept", with respect to their impact on the measured outcome(s). That is, in the absence of any concerns of divergent invalidity or non-random error (that is, assuming that both $\beta_0 = \beta_2 = 0$, Figure 2 would look like Figure 3.

If most of what we observe in q_1 is pure noise, then random measurement error, or unreliability, would indeed be a problem in measuring instructional quality. The consensus, however, is that there is relatively little random error in the measure of q_1, especially if the measure is an index of multiple indicators or is a single item response from multiple classes. There is considerably less agreement about convergent validity. Convergent validity implies that the measure of instructional quality is correlated with other things that are also supposed to be indicators of instructional quality. That is,

multiple indicators of the same underlying concept should be correlated with one another and that the measured SET should be related to measures of actual "learning." The evidence is that student responses about different aspects of teaching are correlated with one other. There is less evidence that measured indicators of instruction on a SET are an endogenous function of actual "learning."[14]

Figure 3: The relation between "true" and measured instructional quality in the presence of random measurement error alone.



The issue of non-random error or divergent (in)validity, is more complicated. Non-random error means that $\beta_0 \neq 0$ and/or $\beta_2 \neq 0$. Thus, the two sources of non-random error are due to mis-estimation of teaching quality because of an intercept error or a slope error. Because any measure of instructional quality is likely to be an ordinal scale at best, so that only relative and not absolute scores are meaningful, the slope error is far more troublesome than intercept error. Thus I consider here only sources of non-random measurement error in q_1 that are due to systematic factors, labeled Z. The overall impact

[14] There is a voluminous literature on the reliability, convergent and divergent validity of SETs. The education literature does not distinguish between convergent validity and reliability. Greenwald (1996, 1997) implicitly equates the two, and provides evidence of both convergent validity and reliability. He also provides evidence of considerable divergent invalidity, arguing that SETs are overlaid with misleading artifacts (or Z-factors, in my language). He does not call this non-random measurement error. However, he does call for the need to adjust for these systematic artifacts. Abrami (2001) is more supportive of the overall validity of SETs. He reviews a large body of education research that supports the reliability, and the convergent and divergent validity of SETs. By contrast, Armstrong (1998) points to the possibility that the correlation between the measured SET and the unmeasured concept might even be negative, suggesting considerable convergent invalidity. Specifically, he argues that SETs might crowd out learning. See also Arthur, Jr. et al. (2003). The economics and management findings tend to rely on formal econometric models for parameter estimation. The education and education psychology literature relies heavily on correlational evidence, which, by definition, cannot detect the issues of exogeneity and the independence of stochastic terms that are critical to establishing both convergent and divergent validity. See for example Threall and Franklin (2001); Kulik (2001); McKeachie (1997); Stapleton and Murkison(2001); Marsh and Roche (1997; Marsh (1987); Ory (2001). Most of these support the overall validity of SETs. By contrast, the economists' contributions, based on parameter estimates that test and adjust for endogenity and use statistical controls, generally cast doubt about both the convergent and divergent reliability of SETs. See for example Becker (2000); Sproule (2002); McPherson (2005); Boex (2000); Krautmann and Sander (1999); Johnson (2003); Nelson and Lynch (1988). The focus in much of the economic and management literature questions divergent validity, focusing especially on the exogenous influence of grades (an outside Z-type variable that does not clearly measure underlying teaching ability) on SETs.

of these systematic influences is represented by β_2, which is assumed to not be zero if there is non-random measurement error. The slope can be either positive or negative. The vector of variables in Z represents the systematic determinants of measured SET scores that are not reflective of underlying learning or teaching effectiveness. These include characteristics of the course (e.g., required or not; introductory or not; size of course; subject of course); characteristics of the instructor (e.g., sex; age; tenured or not); and characteristics of the student (e.g., gender; GPA; major; expected grade in course).[15] These variables may or may not be related to the "true" quality of instruction in the course. In either case, if they remain unmeasured, then the estimated value of q_1 will be biased by the value of β_2 Z. Even though some of the components of Z have a positive influence on q while others have a negative impact, there is no reason to expect that the net influence is zero. Consequently, any measured SET that does not adjust for the influences of the omitted Z-variables will not even be an ordinally accurate estimate of measured instructional quality, q_1.

More generally, the presence of factors (Z) beside "true" instructional quality (q_1') that affect measured instructional quality (q_1) depends not only on the overall magnitude and direction of their impact on q_1, represented by β_2, but also on whether these determinants are related to "true" instructional quality (q_1'), and on whether the relation between "true" and measured instructional quality, represented by β_1 is positive or negative. Two components of the vector of variables represented by Z are of particular importance in this regard, given their presence in the payoff functions of important actors in the typical university. These two components, likely to be positively related to each other, are the expected and actual grade in the course. Grades are clearly valued by students and the provost. If expected and actual grades are also part of the determinants of measured instructional quality, then they will also be valued by faculty, who are rewarded for high instructional quality scores. A considerable body of research, especially in economics, supports the expectation that grades (actual or expected) influence SETs, but this research often suffers from possibly omitted variables, and it rarely distinquishes between the influence of the actual and the expected grade.[16] Most often, the expected grade is studied, rather than the actual grade; but it is the actual grade that "counts" in the payoff functions of the players.

Some (e.g. Marsh and Roche,1997 and Marsh,1987) continue to argue that "true" instructional quality is directly related to higher grades. If that is the case, and if "true" and measured instructional quality are positively related ($\beta_1 > 0$), then the influence of grades on measured instruction ($\beta_2 > 0$) would not threaten the ordinal validity of measured instructional quality. However, even if good instruction leads to higher grades

[15] See for example d'Apollonia and Abrami (1997); Greenwald and Gilmore (1997); Grimes et al. (2004), Langbein, (1994); Stapleton and Murkison (2001); McPherson (2005); Boex (2000); Hoyt and Pallett (1999); and Hamermesh and Parker (2005).

[16] See for example Greenwald (1996), who finds that expected grades influence SETs, as do Greenwald and Gerald (1997); Boex (2000); McPherson (2005); Stapleton and Murkison (2001); Krautmann and Sander (1999); Nelson and Lynch (1989); Zanyenehzadeh (1988); Eizler (2002). Johnson (2003) looks at both actual and expected grades, and Hoyt and Pallett (1999) look at correlations (not parameter estimates) between SETs and actual grades. See also Haskell (1997); Stone (1995); Marsh and Roche (1997); and Marsh (1987). The two papers by Marsh use correlations to argue that there is no direct relation between SETs and grades.

for students, it is not at all clear that good instruction leads to higher scores on measured instruction.[17] Thus, given that we cannot assume that "true" and measured instructional quality are positively related ($\beta_1 > 0$), it is important to examine whether grades influence SETs, independent of other determinants of SETs. If grades influence SETs, then they are as important to the internal political economy of universities as SETs, but grades are not SETs. SETs are an internal currency of contemporary universities, at least in the U.S. and Canada. Outside the university, they have no value.

Grades are different. Grades are signals that have meaning outside the university.[18] If grades become invalid signals, then they will be replaced by other signals, including a demand for expensive graduate education. For their part, rather than use grades as a signal of ability, graduate schools, will be tempted to place more reliance on GRE or LSAT or MCAT scores, which have their own problems of validity.[19] Employers will turn to personal contacts, letters of recommendation, and direct inspection (internships and co-ops) for recruiting. It is not clear that these alternatives are a more informative or less costly signaling device than accurate grades would be (Rosovsky and Hartley, 2002).

While there is considerable research on many of the determinants of SETs, there is limited evidence about the impact of actual grades on SETs.[20] Expected grades are easy to observe in most studies, and most research shows they have a positive influence on measured SETs. But expected grades are not the same as real grades, because they have no direct value to any of the players in the system; what counts is real grades. But it is hard to assemble data that connects SETs and actual grades, because SETs are anonymous at the individual level. Only one study has done this (Johnson, 2003), but it lacks generalizability, because the results are from one year of a special study at Duke

---

[17] For example, d'Appolonia and Abrami (1997) find a moderate positive correlation between measured and objective instructional quality, as do Marsh and Roche (1987) and Hoyt and Pallett (1999); but studies with statistical or experimental controls tend to be less supportive of this conclusion. See for example Rodin and Rodin (1973); Stone (1995); Johnson (2003). As other examples, Yonker and Yonker (2003) find that the grade in accounting II is a negative function of the students' evaluation of teaching in accounting I, with statistical controls for the grade in accounting I, SAT, and GPA. They conclude that high student teaching evaluations are associated with less and not more learning. Greenwald (1996) also provides evidence that rejects the theory that underlying teaching effectiveness accounts for both high grades and high student teaching evaluations.

[18] A large literature supports this contention. See for example Spence (1973, 1974); Stone (1995); Sabot and Wakeman (1991); and Jones and Jackson (1995). I discuss the issue further in the conclusion.

[19] Jesse and Rothstein (2004), using the SAT (not the GRE, which is similar) find that demographic characteristics appear to have more predictive validity of actual college performance than standardized test scores.

[20] See for example Greenwald (1996); Stone (1995); and Haskell (1997), who reviews a large body of evidence. See also Boex (2000), and McPherson (2005); both find that expected grade is a determinant of SETs, but the latter is more convincing because it controls for fixed effects and tests for endogeneity. Nelson and Lynch (1984), Zangenehzadeh (1988), Krautman and Sander (1999) and Murkison and Stapleton (2001) use two (and sometimes three) stage estimates and find that expected grades influence SETs. Studies by Greenwald and Gerald (1997), Eizler (2002) report similar findings. However, Marsh and Roche (1997) and Marsh (1987) find no relation between expected grades and SETs, but their study lacks the statistical and experimental controls that the other studies use. Johnson (2003) is one of few studies to look at both expected and actual grades. Using individual level data in a natural experiment, he finds that both independently raise SET ratings.

University, and suffers from possible omitted variable bias. However, it does show that higher actual (and expected) grades lead to higher SETs.

My study adds to the small body of evidence about the influence of actual and expected grades on SETs. It uses fixed effects for courses and faculty to reduce the threat that uncontrolled factors related to faculty and courses confound the estimate of the influence of actual and expected grades on SETs. It gets around the anonymity problem by using courses rather than students as the unit of analysis. As a result, the data represents all courses (graduate and undergraduate) over a four-year period. Consequently, the results may be more general and less subject to errors due to unobserved heterogenity than those reported by Johnson. Further, the course, rather than the individual student is the appropriate unit of analysis for this study, for individual SETs are not what either the faculty or the provost value. Rather, it is course SETs that determine faculty pay and the provost's assessment of university revenues.

Data

The data for this study come from information on 7686 graduate and undergraduate courses taught at American University from Fall 2000-Spring 2003. All ordinal variables are recorded as averages for the entire class, making them continuous.

Even over the relatively short duration of the data for this study, there is clear evidence of grade inflation at AU, especially in lower level, freshman/sophomore classes. These classes tend to have larger enrollments than other classes; thus, they are likely to be the most profitable for the university. However, freshmen and sophomores are the most likely to transfer, so the informal, unwritten pressure on faculty to help retain these students is clear. One likely consequence is grade inflation, and that is what I observe. Over the period of this study, the mean grade in 100-level courses increases from 3.1 to 3.2; the percent who earn less than a B in these courses drops from 28% to 25%. For 200-level courses, the mean grade increases from 3.1 in Fall '00 to 3.2 in Spring '03, and the percent earning less than a B drops from 29% to 24%. For 300-level courses, the mean grade remains unchanged at 3.3, but the percent below B drops from 22% to 18%, and the median grade increases from B+ to A-. Among higher level classes, no such clear pattern of aggregate grade inflation is apparent.

The central hypothesis of this study is that higher actual grades produce higher SETs. Students react to expected grades in their responses to SETs, because students fill out the SETs before they get their actual grades. However, students are likely to anticipate that expected and actual grades will be highly related to each other, and faculty have little incentive to violate that expectation, as it would damage their reputational capital. To the contrary, faculty often seek to avoid the possibility that a student will "hassle" the professor if the professor gives lower grades than the student expects. Consequently, even though faculty assign grades after students fill out the SET, faculty are ex ante constrained to give grades that usually and approximately correspond to student expectations. Apparently the expectations of both sets of players (students and faculty) are reasonably accurate. While the correlation between expected and actual grade is 0.69 (which means the explained variance is only .47), students only slightly overestimate actual grades. The linear regression coefficient is 0.9 (on a 4-point scale), which means that the expected grade is, on the average, only .1 higher than the actual grade received.

The dependent variable in this study is the SET for the course. The SET score is an average of the overall instructor and course ratings. Each of these is a 1-6 ordinal scale, where 6 is the highest possible score. In this study, the SET is a course average, so it is a fully continuous variable, bounded at 1 and 6. There are two continuous independent variables in the study: the expected and the actual grade. At the individual level, grades are measured on a 4-point ordinal scale. In this study, both are averages for the entire course, so they are continuous measures, bounded at 0 and 4.

Critical to the internal validity of the results of this study are the sets of fixed effects dummy variables. One set of results includes a fixed effect dummy variable for each faculty member. For example, Prof. Michael McCalculus teaches 5 math classes during each of the 3 years in the study; thus the fixed effect dummy for McCalculus is 1 for 15 of the course observations in this vector, and 0 for the remaining 7686-15 = 7671 observations. The same scoring system is used for Ann Amlit. Prof. Amlit teaches only 4 courses each year, so her dummy vector consists of 12 "1's" and 7686-12 = 7674 "0's". Few courses are jointly taught, so there is little overlap between faculty dummy variables. There are about 1200 individual faculty (tenured, tenure track, adjuncts, etc.), each with a separate dummy vector. Including in the regression a vector for each faculty member controls not only for easy-to-measure differences among faculty (age, race, gender, experience, subject) but also for harder-to-measure differences (expressiveness, looks) and for un-measurable and idiosyncratic differences among faculty. The set of faculty dummies also implicitly controls for many course-related differences among faculty. For example, some faculty teach only graduate courses; other faculty teach only relatively large courses.

Another set of results includes a fixed effect for each course, without regard for who taught the course. So if Math101 is taught 3 times every semester during the 3 years of the study, the Math101 course dummy is coded “1” for 18 of the 7686 observations in the vector; the remaining course observations in the vector are coded “0”. There are about 1300 course dummies, one for each unique course number in the data. These dummies control for differences among courses, no matter whether they are measurable (e.g. required; size of course; topic; average ability level of students in the course) or hard to measure or not measurable (e.g., average level of interest of students in the course; condition of room; whether the laboratory equipment for the course malfunctions; whether the TA speaks English).

The final set of results includes fixed effects for both sets of dummies. There is collinearity between some faculty and course dummies when the faculty and the courses are not separable. However, since the separate dummy variables are not of inherent interest (except maybe to the provost), that the analysis drops the collinear dummies is no problem. The advantage of this regression, with some 2200 fixed effect dummies and close to 5400 remaining degrees of freedom, is that it controls best for the entire set of potentially omitted variables. It reports parameter estimates for actual and expected grade that are least likely to suffer from omitted variable bias.

I also test the estimates for simultaneous equation bias. If, as some allege, high actual (and expected) grades are a function of good teaching, presumably measured by high SETs, then it suggests that actual and expected grades are endogenous to SETs. In may also be the case that courses with high SETs crowd out learning. In either case, the results in the OLS fixed effect models would be biased. I report the results from this test.

Results

Table 1 reports the means and standard deviations of the continuous variables in the study. The average of the SET score is 4.9 on a 6-point scale. On the scale, a score of "6" is "superior;" a score of 5 is "very good." This illustrates the possibility of a Lake Woebegone effect, where all (or most) of the teachers are "above average," that is, better than "good", which is an SET score of 4. In fact, the median score is 5, meaning that half the teachers are "very good" or better, and half are "not very good." Thus, given the common practice of ranking instructors for the purpose of using SETs to allocate merit pay, 50% of teachers are "not very good."

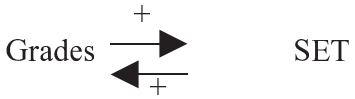
Faculty reciprocate the high regard with which students rate them with average grades that are better than "B". The average grade is 3.33 on a 4.0 scale; the median grade is 3.36. Confirming the previous observation, based on bivariate analysis, about the relative accuracy of student grade expectations, the univariate results in Table 1 show that students expect a grade that is close to what they receive. The average expected grade is 3.45, only 0.1 point higher than the average student actually receives. It is apparent that the range of actual grades is considerably higher than the range of expected grades. There is at least one case where the actual average grade in the entire class is D, or "1." There is also at least one case in which the actual average grade in the entire class is A, or 4.0. The only way that the average for a class can be 4.0 is for everyone in the class to receive an A. (That is not the case for an average of D, because a failing grade is "0.") The highest average expected grade is an A; the lowest is only 1.88, about a C-. There appears to be more optimism by students about their likely grade at the low end, where consistent disappointment may result in pressure on faculty to raise actual grades.

Table 2 reports the OLS regression results, with robust standard errors. The model fit is high, because of all of the fixed effects controls. The R^2 is over 0.98 in all 3 of the models, and the root mean square error is low. Yet the model is not over-determined. Depending on the number of fixed effect (FE) controls, the model has from 5391 to 6514 remaining degrees of freedom. The results are consistent no matter whether fixed effects for courses, faculty, or both are held constant. Even with this extensive number of controls for each course, instructor or both, both the actual and expected grade have a significant positive effect on the average SET that instructors receive. The expected grade consistently has a larger effect on the SET than the actual grade. This makes sense because the students report the expected grade at the same time they report their evaluation of the instructor. The estimated magnitude of the impact of each declines as the number of fixed effect dummy controls increases. It is therefore likely that the estimates least likely to reflect omitted variable bias are those in the third panel of Table 2. Those estimates reveal that the impact of a unit increase in the expected grade (say, from B to A, which is where the bulk of the observations lies) would raise the instructor's rating by an average of nearly 0.5 on a 6 point scale. In a rank order system, this is not irrelevant; on a percentage basis, each additional .5 of a point is an 8% higher ranking. Over time, the effect of an additional 0.5 in the SET on an increase in a merit pay ranking could be considerable. The incentive is for faculty to grade leniently during the semester. Students then expect lenient grades, and reciprocate by ranking the instructor generously on the SET. Faculty play the final round in this game by awarding an actual grade that is, on the average, close to what students expect (only 0.1 less, on a 4 point scale). Unsurprisingly, controlling for the expected grade, actual grades have less of an impact on the SETs than expected grades. Specifically, using the results in the third panel of

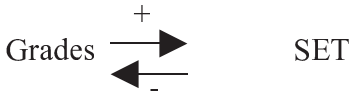
Table 2, a one point increase in the average actual grade (say, from B to A, which is also where the observations lie) raises the SET by less than 0.1 point on a 6 point scale.[21]

Results: reprise

It is important that these estimates not only reduce the possibility of omitted variable bias by including numerous fixed effects that proxy faculty and course characteristics (and student characteristics that are unique to each course). It is also important that they reflect the exogenous impact of grade on SET, if they are to support the maintained hypothesis in this study, which is that faculty grade leniently in order to get higher ratings from students. There are two opposing counterclaims regarding the assumed exogeneity of actual grades on SETs. The first is the contention that SETs are generally a valid indicator of good teaching and more learning (Marsh, 1987; Marsh and Roche, 1997). If this claim were true, classes with high SETs will also have students who got good grades because they learn more. This is a claim that the relation between average grades and SETs in courses looks like this:



The second, opposing, claim is that satisfied students actually learn less, which means that they wind up with lower grades (Rodin and Rodin, 1973; Johnson, 2003; Yonker and Yonker, 2003). This is a claim that the relation between average grades and SETs in courses looks like this:

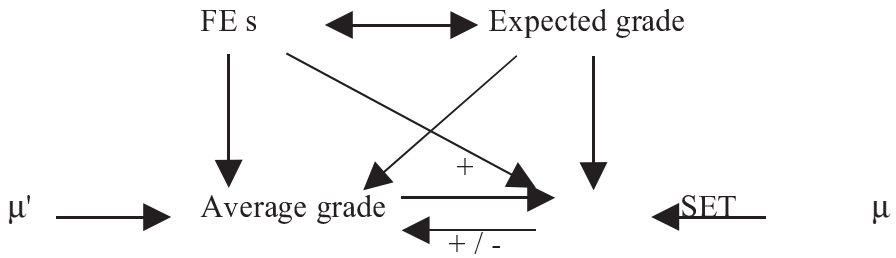


To examine these claims, it is necessary to test for the exogeneity of average grades in the model estimates reported in Table 2. Both claims imply that the average grade is not exogenous, but they disagree about the direction of the reverse causation. I use a Hausman test of simultaneity (Gujarati, 2003: 754-756). The test requires assuming that some variables in the model are exogenous. I assume that the course and

[21] These results are robust to various methods of estimation. For example, the dependent variable is not truly continuous, since it is bounded at the bottom by 1 and at the top by 6. To adjust for this, I used Tobit to re-estimate the model parameters. Using Tobit estimates with nearly full fixed effects does not change the results at all. Unfortunately, using the full set of fixed effects made it impossible for the Tobit estimates to converge. This is likely because, with full fixed effects, the difference between expected and actual values of the dependent variable is so close that the estimates cannot get "worse" after many trials, so the maximum likelihood estimator cannot find a maximum. In effect, the derivative that tests for a maximum never turns negative. Removing some of the fixed effects finally produced a model that converged. In that model, the Tobit ML estimates and hypothesis tests did not differ from those reported. I also estimated the data separately for graduate and undergraduate courses. The results for the undergraduate model do not differ from those reported. The results for the graduate model also do not differ substantially in sign or significance of the parameter estimates. Specifically, the parameter estimate for impact of actual grade on SET is slightly larger than that reported in Table 2 ($b = .18$, $p > .02$), while the estimate for the impact of expected grade is slightly smaller ($b = .28$, $p > .02$). The goodness of fit statistics for both models resemble those reported in the 3rd column of Table 2.

faculty fixed effect (FE) dummies and the expected grade are exogenous to both the average grade and the SET. In this formulation, the FE dummies are, indeed, fixed; and the expected grade is an exogenous predictor of the average grade, which is received subsequent to the report of the expected grade. The issue concerns whether average grade is or is not endogenous to the SET; and, if it is endogenous, whether the reverse causation is positive or negative.

The diagram below represents the structural set-up of the issue:



The test proceeds by separating the exogenous component of the observed average grade from the rest of it. Assuming that the FE dummies and the expected grade are truly exogenous to average grade, then the predicted value of average grade must be exogenous, because it is a linear function of the exogenous FE dummies and expected grade. The average grade residual contains other causes of average grade, some of which may be due to the endogeneity of the SET. In fact, the residual is quite small. The FE dummies and the expected grade account for 99.46% of the average grade.[22] The remaining component of the observed average grade (the residual) could either be noise (μ') or it could represent the component of average grade that is affected by the SET. If the residual were noise alone, and represented no endogenous component, then the average grade residual would have no direct impact on the SET. In that case, in a regression of the SET on the predicted (or actual) average grade, expected grade, the FE dummies, and the average-grade-residual, the residual term would have no significant impact. On the other hand, if the residual included the endogenous component, as well as noise, its impact in this regression would be significant. Given that the FE dummies and the expected grade explain most of the variance in observed average grade, it seems reasonable to assume that the residual is not noise, but rather represents a fairly clear test of endogeneity. Further, if the residual coefficient represents endogeneity, then its sign is indicative of the direction of the endogenous influence. If the sign is positive, that would be consistent with the argument that high SETs reflect good teaching, as they bring about more learning and higher grades.[23] If the sign of a significant residual is negative, it would signify that high SETs may crowd out learning.

[22] After controlling for the FE dummies, the expected grade has an important and statistically significant impact on average grades. The partial regression coefficient is over 0.5, meaning that, controlling for the FE dummies, for every grade point increase that is expected, students receive about half of what they expect.

[23] An alternative explanation is that the positive residual indicates that reciprocity between high SETs for faculty and higher grades for students is in full effect.

Table 3 reports the results. First, the results in this table merely rearrange the same information in the regression reported in the third column of Table 2, so the goodness of fit statistics are identical. Second, and most important, the results "purge" the measure of the average actual class grade from expected endogeneity. There are two measures of the actual grade, consistent with the tests in Gujarati (2003: 754-756). One is the observed average actual grade; it contains the noise and the possible endogeneity that may be present in the model estimates of Table 2, but in Table 3, the residual (which is the difference between the predicted and observed average grade) statistically controls for the possible endogeneity. The other measure is the average grade predicted by the FEs and the expected grade. Theoretically, it is not clear which of these should be used. If the observed average grade is really a composite only of the exogenous variables (the FEs and expected grade) and the endogenous effect of the SETs, then the model with the predicted rather than the observed average grade would be preferable. But this is a strong assumption. In either case, however, the coefficient for the actual average class grade remains positive, significant, and larger in magnitude than the estimate reported in column 3 of Table 2. The results strengthen the inference that giving high grades raises SETs. Specifically, for each additional average grade on a 4-point scale, SETs go up by 0.9 on a 6-point scale. Interestingly, when possible endogeneity is removed, the variable that loses its steam is expected grade, which is no longer significant in either model reported in Table 3. Once possible endogeneity of the average grade is accounted for, the expected grade has no direct effect on how students rate faculty.[24]

While the results in Table 3 clearly uphold the conclusion that faculty are rewarded with higher SETs if they reward students with higher grades, the sign of residual variable depends on the specification of the endogeneity test. Under one specification, the sign of the residual is negative; under the other, it is positive. Consequently, the results give no clear indication about whether some component of the SET is a measure of "good" teaching and more learning, or easy class content and less learning. Under the assumptions of the model, if the residual of average grade contains little or no noise, but reflects the endogenous impact of the SET on the average grade, if the SET raises the average grade, it would be a sign that the SET results in more learning. The residual would be significant and positive. When the predicted average grade is used as the measure of the main theoretical variable in the study, the residual is indeed positive and significant. (See column 2 of Table 3.) On the other hand, when the observed average grade is used as the main theoretical variable, the residual is negative and significant. I suspect that these different results reflect in part the impact of expected grade on both the SETs and the average grades. Recall that faculty give students close to the grade they expect, but the actual grade is a bit lower than what students hope for. Faculty give in to student expectations in order to get a higher SET….but they are not complete suckers in this game. However, it is dangerous to over-interpret the coefficient on the residual variable. The central point is that, even when possible endogeneity is accounted for, the structural impact of higher grades on student evaluations of faculty (the SET) remains significant and positive. This study was not designed to provide a conclusive test regarding the structural impact of SETs on grades, and the results reported

[24] While this is speculative, the expected grade may have an indirect effect. Student expectations about their grade, which reflect actual pre-final grades, constrain faculty to (roughly) conform to those expectations, with only a little disappointment in store for the student. Hence the actual grade reflects the expected grade, and also elicits higher SETs.

are not consistent with one single interpretation. By contrast, the results regarding the impact of grade on SETs are consistent: faculty who give higher grades get higher SETs, controlling for unobserved heterogeneity among courses (and the students in them) and faculty.

Conclusion

It appears likely that the current system will be retained.[25] It is valued by students and by the provost. Students like to be satisfied, and they like higher grades. The provost wants to retain students, and prefers a low cost system to monitor faculty that looks "objective." (See Becker, 2000; Sproule, 2000.) What about the faculty? Why should they value this system? Given the use of rank order merit pay systems based on SET scores, the median faculty is "above average," creating a Lake Woebegon effect. Because 50% are above the median raise, it will probably survive a vote. Of course, 50% are below the "average," which is really a median. This group is likely to be demoralized. The result is driven by rank order evaluations, which are an important component in many merit pay/continuous reward systems.

Yet many scholars note problems with relative pay systems. For example, McKeachie (nd), argues that, even if SETs are valid and reliable, when 90% of teachers at the University of Michigan are rated "excellent", but, with a relative pay system, 50% are still below the median rating, the consequence is de-motivating and demoralization. If this demoralization has observable consequences for quality teaching, it would serve as another warning that extrinsic rewards, especially when they are awarded on a piece-rate basis, can drive out intrinsic motivations, resulting in a costly loss of performance (Frey and Oberholzer-Gee, 1997; Kreps, 1997). McKeachie further argues that the SET should not be used as a continuous rating scale, where 3.1 > 3.0 and insignificant decimals determine pay raises. Rather, they should be used as a discrete standard that should be met (within a certain confidence interval and level). Kane and Staiger (2002) note that variances of rating systems are higher when there are small classes or fewer courses, and that a few disgruntled students can have a large impact on averages.[26]

The Lake Woebegon effect is also true for students: all students are (well) above average (=C). This has come about since colleges and universities started treating students as consumers, shifting the focus of managers to customer satisfaction rather than education.

Why should we worry about grade inflation? It appears to be the case that grades have inflated, and that grade inflation matters because grades are external signals.

---

[25] Crumbley and Fliedner (2002) report that accounting department administrators realize the biases of SETs, but do not want to replace or fix them.

[26] There is a large literature on problems with relative pay systems. McKeachie (nd) and England (nd), in addition to urging the use of discrete cutpoints on the SET also suggest using student outcomes to evaluate instructional quality. Baker (1992), Gibbons and Murphy (1990), and Lazear (1989) all point to problems with relative pay systems, which is what the merit pay system of more pay for higher SETs is. Such a system may be nonoptimal if there is a possibility that employee A can sabotage the output of employee B. While this is not directly possible in university teaching, the use of merit pay based on relative performance does not encourage cooperation among peers in the same department. Drago and Garvey (1998) find that employees paid under a piece-rate system, so that their pay is a positive linear function of some performance measure, cooperate less than similar employees paid under a discrete pay system. More important in a university setting is that relative pay does not encourage faculty members in the same department to hire someone who can teach "better" (as measured by the SETs), or do more research, than they do. With merit pay based on relative performance, no one wants a "rate-buster."

Rosovsky and Hartley (2002) note that grades have gone up but that SAT scores have gone down. Grades have inflated since the 1970s, but they leveled off in the 1990's because they are constrained at the top. Rosovsky and Hartley equate the beginning of the increase in grades and drop in SATs with the introduction of SETs. Sabot and Wakeman-Linn (1991) also provide evidence of grade inflation, as does the Wall Street Journal (2003). Grades have inflated in Canada too. Anglin and Meng (2000) note that, from 1973 to 1993, grades went up, especially in English, Biology, and general Arts and Sciences areas; there was not much of an increase in math grades over this period. Sabot and Wakeman-Linn point further to the social consequences of grade inflation; since there is less grade inflation in math and science courses, fewer kids major in science and math. This choice is rational for students, since higher grades lead to higher pay later on (Jones and Jackson, 1990).

But higher grades, if they are not a reliable signal of performance, only exacerbate the problem of monitoring in the employment market. Shapiro and Stiglitz (1984) develop a model of unemployment as a worker discipline device that results from the information asymmetry between principals (buyers of labor) and agents in employment markets. They argue that grade inflation may exacerbate this asymmetry, and raise monitoring costs even more. The consequence would be a higher unemployment equilibrium that would not be socially optimal. Spence (1973, 1974) contends that grade inflation reduces the negative correlation between education costs and productivity that is necessary to make education an efficient signal. In other words, even though the costs of education increase at the margin, productivity does not. One possible consequence of grade inflation is an increased demand for higher degrees, such as the MA; but this only raises costs, not productivity. Stone (1995) notes the mismatch of grades and actual student achievement from 1963 to 1980, when SATs declined and grades went up. Since then, he says, the pattern has not reversed itself. An efficiency implication is that, since college instruction is more expensive than high school instruction, the increased cost of college, with no apparent increase in effectiveness, may represent misdirected expenses that might more profitably be spent on high school. Stone also joins others in bemoaning the impact of SETs on college instructors' loss of moral authority. Sabot and Wakeman (1991) point out that grades are a signal to students of their comparative advantage in skills; grade inflation biases the signal, causing students to choose a course of study that may not be optimal. Jones and Jackson (1995) find a positive correlation between grades and earnings, implying that grades continue to serve as external signals in the labor market, even if they may be biased if not noisy.

One policy implication of these results is that it may be possible to reduce the connection between SETs and grades by avoiding the use of a continuous reward system between SETs and faculty raises. Under a continuous reward system, the relation between faculty pay and SET, holding research and service constant, is:

$$P_t = \alpha_1 \; q_1 + \gamma P_{t-1}$$

where q_1 is measured teaching ability. By contrast, under a discrete reward system, the relation between the faculty pay and measured teaching becomes:

$$P_t = \alpha'_1 \; q_1 \; (q_1 > q_{1\,min}) + \gamma P_{t-1}.$$

This reward system is no longer purely a rank order system. Faculty (perhaps with administrators) would collectively decide on an acceptable SET rating, as well as a confidence level and confidence interval around the acceptable score. Faculty who are (statistically) above the acceptable rating get a raise (or tenure, or promotion), others do not.

Many other policy suggestions exist for either reducing the influence of SETs on faculty pay or for improving the validity or reliability of measured SETs. To improve the reliability of SETs, one suggestion is to remove outliers. To improve the validity of SETs, another suggestion is to measure actual improvement in student learning by a pre- and post-test ("value added") and to use that along with or instead of the usual SET. Another suggestion for improving the validity of SETs is to weight them, formally or informally, by factors such as actual grade or other variables, such as whether the class is required or uses math, which affect the measured value of SETs. To reduce the influence of SETs on grades, one approach does not actually change the use of SETs, but instead changes the use or distribution of grades. For example, it is possible to constrain grades in various ways, such as permitting no more than some percent of a class to receive A's. The percentage could easily vary by level of class and by topic area. The problem with these approaches is that none of them affect the use of SETs as a rank order, piece rate system to measure faculty teaching performance. Unless one can devise a valid and reliable measure of teaching, which is inherently hard to measure, piece rate rewards are likely to have adverse effects on the very performance outcome that all of the players in the system would really prefer to enhance.

More generally, the results in this study suggest that efforts by public and non-profit agencies, and the federal government in particular, to increase productivity by using piece-rate pay systems to reward good performance may have unintended adverse consequences. Piece-rate pay systems make raises linearly dependent on measured performance such that if person i's measured performance is better than j's, then, all else constant, i gets a bigger raise than j. Measuring good college or university teaching is hard enough; even then, rewarding it with higher pay appears to encourage higher grades, but not more learning. But measuring good performance of employees in many public agencies is even more difficult. This is especially the case when the agency's output is hard to measure and collectively consumed. For example, it is hard to measure "national security" in the case of the Defense and Homeland Security Departments. It is hard to measure "domestic security" in the case of criminal law enforcement agencies, "justice" in the case of civil law enforcement agencies, and "national interests" in the case of foreign affairs agencies, especially those like the State Department not directly connected to trade and commerce. The process is complicated even more when the employee's own output is hard to measure, when the agency output is the product of the joint effort in individual employees, and when the connection between individual employee actions and agency output is complicated or totally unknown (Langbein and Jorstad, 2004).

Based on the results in this study, and others, employees will do what they are paid to do, regardless of the impact of their actions on the collective mission of the organization. This is productivity enhancing when employee output is easy to measure and when the relation between employee and organization output is known. But when both employee and organization output are hard to measure and when the production function linking them is unknown, rewarding employee performance by a piece-rate pay system is likely to exacerbate the information asymmetry between higher management

and "street" or "bench" level employees. With piece rate pay, employees have little incentive to question managerial directives, as expressed in the pay system. For instance, in the case of university SETs, managers reward employees for higher SETs, not for "better teaching." The consequence is more managerial control, but not necessarily improved organizational performance.[27] Comparing and paying university faculty for higher SETs appears to encourage inflated grades. In other settings, particularly where employee and organizational outputs are hard to measure and collectively produced and consumed, the adverse consequences of piece rate pay could be far worse.

An alternative is to rely more on open contracts and a discrete pay system. Such systems rely on shared norms and horizontal communications among peers who monitor one another and collectively carry out the work of the group. In this system, peers sanction shirkers and free-riders who then correct their ways, or they banish or excommunicate non-cooperators or move them to a location where they do less harm (Chen, 2000; Frank, 1998; Gibbons, 1998; Kreps, 1990, 1997; Miller, 2000, 2001; Miller and Whitford, 2002; Schotter, 1998).[28]

While this study does not directly compare discrete to piece-rate pay systems, it does provide evidence that piece-rate pay-for-performance, in a context where it might arguably even "work," has adverse unintended consequences. Piece-rate pay-for-performance works best in enterprises that produce measurable output and where the connection between individual and collective output is clear and known (Baker, 1992). Importing such a system to other enterprises, especially public sector agencies whose employees collectively produce outcomes that are collectively consumed, is likely to make these agencies less and not more socially productive.

[27] Even in the case of public education, where the goal of improving academic achievement is relatively measurable using standardized tests, the effectiveness of pay-for-performance is at best marginal. For example, Dee and Keys (2004), exploiting the random assignment of teachers and students to classes in the Tennessee STAR (Student Teacher Achievement Ratio) experiment, find that teachers working under a merit pay system are only marginally more effective than statistically equivalent others who are not subject to merit pay. Inexplicably, the results were more consistent for gains in math than for gains in reading. Notably, the marginally but nonetheless apparently effective merit pay system in this study was discrete, and not a piece-rate pay system.

[28] In the context of "lower" education, Akerlof and Kranton (2002) develop a formal model of a hypothetical elementary or high school that incorporates the possibility of repeated interactions among students and between students and teachers. The results show how the potential importance of shared norms and cooperative reinforcement (and punishment) can result in better student (and teacher) performance. They point to empirical evidence that is consistent with the predictions of their theory.

Table 1: Descriptive statistics: actual grade, expected grade, and SET

	Mean	Std. dev.	Min.	Max.	Median	Nobs.
Actgrade	3.33	0.41	1	4	3.36	7686
Expgrade	3.45	0.31	1.88	4	3.47	7686
SET	4.89	0.68	1.13	6	5	7686

Table 2: Regression of SET on actual and expected grade, fixed effects (robust standard errors; no intercept).

	Unstd. est.	t-stat.	Unstd. est.	t-stat.	Unstd. est.	t-stat.
Actgrade	0.17	5.42	0.16	5.89	0.11	3.13
Expgrade	0.92	18.32	0.86	17.69	0.57	9.53
Faculty FE	NO		YES		YES	
Course FE	YES		NO		YES	
N obs.		7686		7686		7686
R-squared	.9880		.9918		.9945	
F-value	433		672		442	
Root MSE	.58		.48		.44	

Table 3: Endogeneity test. Regression of SET on actual (and predicted) grade, expected grade, faculty and course dummies (not shown), and expected-grade-residual (robust standard errors; no intercept).

	Unstd. est.	t-stat.	Unstd. est.	t-stat.
Actgrade	0.94	5.47		
Actgrdhat			0.94	5.47
Expgrade	0.14	0.99	0.14	0.99
Residual	-0.83	-4.53	0.11	3.13
Faculty FE	YES		YES	
Course FE	YES		YES	
N obs.		7686		7686

List of references

Philip C. Abrami (2001), "Improving Judgments About Teaching Effectiveness Using Teaching Rating Forms," New Directions for Institutional Research 108 (Spring): 59-87.

George A. Akerlof and Rachel E. Kranton (2002). "Identity and Schooling: Some Lessons for the Economics of Education." Journal of Economic Literature 40 (4), December: 1167-1201.

P.M. Anglin and R. Meng (2000). "Evidence on Grades and Grade Inflation at Ontario's Universities, " Canadian Public Policy Analysis-Analyse de Politiques 26(3), September: 361-368.

J. Scott Armstrong (1998). "Are Student Ratings of Instruction Useful?" American Psychologist 53(Nov.): 1223-24.

Winfred Arthur, Jr., Travis Tubre, Don S. Paul, Pamela S. Edens (2003). "Teaching Effectiveness: The Relationship Retween Reaction and Learning Evaluation Criteria." Educational Psychology 23(3), June: 275-85.

George P. Baker (1992). "Incentive Contracts and Performance Measurement." Journal of Political Economy 100(3): 598-614.

William Becker (2000). "Teaching Economics in the 21st Century." Journal of Economic Perspectives 14(1): 109-119.

L. F. Jameson Boex (2000). "Attributes of Effective Economics Instructors: An Analysis of Student Evaluations." Journal of Economic Education (Summer): 211-227.

Richard T. Boylan (2004). "Salaries, Turnover and performance in the Federal Criminal Justice System." Journal of Law and Economics 47 (April): 79-91.

John Brehm and Scott Gates (1997). Working, Shirking and Sabotage: Bureaucratic Response to a Democratic Public. Ann Arbor: University of Michigan Press.

Dave Buck (1998). "Student Evaluations of Teaching Effectiveness Measure the Intervention, not the Effect." American Psychologist 53(11), Nov.:1224-26 53.

Jongmin Chen (2000). "Promises, Trust and Contracts." Journal of Law, Economics and Organization 16(1) April: 209-232

D. L. Crumbley and E. Fliedner (2002). "Accounting Administrators' Perceptions of SET Information," Quality Assurance in Education 10 (4): 213-222

S. d'Apollonia and P. C. Abrami (1997), "Navigating Student Ratings of Instruction." American Psychologist 52(11): 1198-1208.

Thomas S. Dee and Benjamin J. Keys (2004). "Does Merit Pay Reward Good Teachers: Evidence from a Randomized Experiment." Journal of Public Policy Analysis and Management 23 (3): 471-488.

Robert Drago and Gerald T. Garvey (1998). "Incentives for Helping on the Job: Theory and Evidence," Journal of Labor Economics 16(1), January: 1-25.

C. F. Eizler (2002). "College Students' Evaluations of Teaching and Grade Inflation," Research in Higher Education 43(4): 483-501.

J. England (nd). "How Evaluations of Teaching Are Used in Personnel Decisions." The Professional Evaluation of Teaching. American Council of Learned Societies Occasional Paper No 33 http://www.acls.org/op33.htm

Robert Frank (1991). "Social Forces in the Workplace." Pp. 181-202 in Kenneth J. Koford and Jeffrey B. Millers, eds. Social Norms and Economic Institutions. Ann Arbor: University of Michigan Press.

Bruno S. Frey (1993). "Does Monitoring Increase Work Effort? The Rivalry with Trust and Loyalty," Economic Inquiry 31 (Oct.): 663-670.

Bruno S. Frey and Felix Oberholzer-Gee (1997). "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-out" American Economic Review 87(4): 746-755.

Robert Gibbons (1998). "Incentives in Organizations," Journal of Economic Perspectives 12(4) Fall: 115-132.

Robert Gibbons and Kevin J. Morphy (1990). "Relative Performance Evaluation for Chief Executive Officers," Industrial and Labor Relations Review 43 (suppl., Feb.): 305-515.

Mary Gray and Barbara R. Bergman (2003). "Student Teaching Evaluations: Inaccurate, Demeaning, Misused," Academe (Sept.-Oct.): 44-46

A. G. Greenwald (1996). "Applying Social Psychology to Reveal a Major (but Correctable) Flaw in Student Evaluations of Teaching," (Mar. 1) http://weber.u.washington.edu/Nagg/asptramf.htm

A. G. Greenwald (1997). "Validity Concerns and Usefulness of Student Ratings of Instruction," American Psychologist 52(11): 1182-1186.

A. G. Greenwald and M. Gerald (1997). "Grading Leniency is a Removable Contaminant of Student Ratings," American Psychologist 52 (11): 1209-1217.

A. G. Greenwald and J. M. Gillmore (1997). "No Pain, No Gain? The Importance of Measuring Student Workload in Student Ratings of Instruction. " Journal of Educational Psychology (89): 743-751.

Paul W. Grimes, Meghan J. Millea and Thomas W. Woodroff (2004). "Grades -Who's to Blame? Student Evaluation of Teaching and Locus of Control," Journal of Economic Educucation 35 (2) Spring: 129-147.

Damodar Gujarati, Basic Econometrics, 4th ed. McGraw-Hill. New York. 2003.

D. Hamermesh and A. Parker (2005). "Beauty in the Classroom: Professors' Pulchritude and Putatative Pedogogical Productivity," Economics of Education Review (forthcoming) and NBER Working Paper 9853.

Richard Hartwig (2004). "A Tiny Ring of Power: The Department Chair and Golden Rule Management." Journal of Public Affairs Education 10 (1): 31-42.

Robert E. Haskell (1997). "Academic Freedom, Tenure and Student Evaluation of Faculty: Galloping Polls in the 21st Ccentury." Education Policy Analysis Archives 5(6): Feb. 12.

Bengt Holmstrom (1979). "Moral Hazard and Observability," Bell Journal of Economics 10 (Spring): 79-91

Bengt Holmstrom (1982). "Moral Hazard in Teams" Bell Journal of Economics 13(2) Autumn: 324-40.

D. P. Hoyt and W. H. Pallett (1999). "Appraising Teaching Effectiveness: Beyond Student Ratings," Idea Paper #36 (Nov.). IDEA Center, Kansas State Univ., Manhattan, Kansas.

Paul Isely and Halinder Singh (forthcoming). "Do Higher Grades Lead to Favorable Evaluations?" J. of Economics of Education.

J. M. Jesse and M. Rothstein (2004). "College Performance Predictions and the SAT," Journal of Econometrics 121 (1-2), July-Aug.: 297-317.

Valen E. Johnson (2003). Grade Inflation: A Crisis in College Educucation. New York: Springer.

Ethel B. Jones and John D. Jackson (1999). "College Grades and Labor Market Rewards," Journal of Human Resources 25(2), Spring: 253-

K. Kanagaretnam, R. Mathieu, and A. Thevaranjan (2003). "An Economic Analysis of the Use of Student Evaluations: Implications for Universities." Managerial and Decision Economics 24(1), Jan.-Feb.: 1-13.

Thomas J. Kane and Douglas O. Staiger (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures," Journal of Economic Perspectives 16 (4), Fall: 91-114.

Steven Kerr (1975). "On the Folly of Rewarding for A While Hoping for B," Academy of Management Journal 18(4) Dec.: 769-83.

James A. Kulik (2001). "Student Ratings: Validity, Utility and Controversy." New Directions for Institutional Research 27(5): 9-25

Arthur Krautmann and W. Sander (1999). "Grades and Student Evaluations of Teachers." Economics of Education Review 18: 49-53.

D. M. Kreps (1990). "Corporate Culture and Economic Theory." In James E. Alt and Kenneth Shepsle, eds., Perspectives on Positive Political Economy. New York: Cambridge University Press.

D. M. Kreps (1997). "Intrinsic Motivation and Extrinsic Incentives." American Economic Review 87(2): 359-364

James A Kulik, (2001) "Student Ratings: Validity, Utility & Controversy." New Directions in Institutional Research, 27(5): 9-25.

Laura I. Langbein (1994). "The Validity of Student Evaluations of Teaching," PS: Political Science and Politics," Sept: 545-552.

Laura Langbein and Connie Jorstad (2004). "Productivity in the Workplace: Cops, Culture, Communication, Cooperation, and Collusion." Political Research Quarterly 57 (1), March: 65-79.

Laura Langbein and Kevin Snider (1999). "The Impact of Teaching on Retention: Some Quantitative Evidence," Social Science Quarterly 80(3): 457-72.

Edward P. Lazear (1989). "Pay Equality and Industrial Politics," Journal of Political Economy 97 (June): 561-580.

Edward P. Lazear (2000). "The Power of Incentives," American Economic Review (Papers and Proceedings) 90 (2), May: 410-414.

Henry M. Levin (1991). "Raising Productivity in Higher Education." Journal of Higher Education, 62(3), May-June: 241-262.

H.W. Marsh (1987). "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research." International Journal of Educational Research 11: 253-388.

Herbert W Marsh & Lawrence A Roche (1997). "Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias and Utility," American Psychologist 52(11) 1187-1197.

W. H. McKeachie (nd). "Student Ratings of Teaching," in The Professional Evaluation of Teaching, American Council of Learned Societies Occasional Paper No. 33 http://www.acls.org/op33.htm

W. H. Mckeachie (1997). "Student Ratings: The Validity of Use," American Psychologist 52 (11): 1218-1225

Michael A. McPherson (2005). "Determinants of How Students Evaluate Teachers. " Journal of Economic Education 36 (4), Fall (forthcoming).

Gary Miller (2000). "Above Politics: Credible Commitment and Efficiency in the Design of Public Agencies. Journal of Public Administration Research and Theory 10: 289-328.

Gary Miller (2001). "Why is Trust Necessary in Organizations? The Moral Hazard of Profit Maximization." Ch. 10 in Karen S. Cook, ed., Trust in Society. New York: Russell Sage Foundation.

Gary Miller and Andrew Whitford (2002). "Trust and Incentives in Principal-Agent Negotiations: The 'Insurance-Incentive' Tradeoff." Journal of Theoretical Politics 14 (April): 231-267.

D. Nagin, J. Rebitzer, S. Sanders, and L. Taylor (2002). "Monitoring, Motivation and Management: The Determinants of Opportunistic Behavior in a Field Experiment." NBER Working Paper 8811 (March).

J. P. Nelson and K. A. Lynch (1984). "Grade Inflation, Real Income, Simultaneity and Teaching Evaluations." Journal of Economic Education 15 (Winter): 21-37.

J. P. Nelson, K. A. Lynch and Hamid Zanyenehzadeh(1988). "Grade Inflation: A Way Out." Journal of Economic Education 19(3), Summer: 217-226.

Chezy Ofir and Itamar Simonson (2001), "In Search of Negative Customer Feedback: The Effect of Expecting to Evaluate on Satisfaction Evaluations," Journal of Marketing Research 38: 170-182.

J. C. Ory (2001). "Faculty Thoughts and Concerns About Student Ratings," New Directions for Teaching and Learning 87 (Fall): 3-15.

J. C. Ory and K. Ryan (2001). "How do student evaluations measure up to a new validity framework?" New Directions for Institutional Research 27(5): 27-44.

Howard Polio and Hall P. Beck (2000) "When the Tail Wags the Dog: Perceptions of Learning and Grade Orientation In, and By, Contemporary College Students and Faculty." Journal of Higher Education 71 (1), Jan/Feb: 84-102.

Richard E, Redding (1998). "Students' Evaluations of Teaching Fuel Grade Inflation," American Psychologist 53 (11), November: 1227-8.

Miriam Rodin and Burton Rodin (1973). "Student Evaluations of Teachers." Journal of Economic Education 5 (Fall): 5-9

Henry Rosovsky and Matthew Hartley (2002). Evaluation and the Academy - Are We Doing the Right Thing: Grade Inflation and Letters of Recommendation. Cambridge, Mass.: American Academy of Arts and Sciences.

Richard Sabot and John Wakeman-Linn (1991). "Grade Inflation and Course Choice," Journal of Economic Perspectives 5(1), Winter: 159-170.

Andrew Schotter (1998). "Worker Trust, System Vulnerability and the Performance of Work Groups." Ch. 14 in Avner Ben-Ner and Louis Putterman, Economics, Values and Organization. New York: Cambridge University Press.

P. Seldin (1993a). "How Colleges Evaluate Professors: 1983 versus 1993." American Academy of Higher Education Bulletin (Oct.): 6-8, 12.

P. Seldin, ed, (1999), Changing Practices in Evaluating Teaching. Bolton, Mass.: Anker.

Carl Shapiro and Joseph Stiglitz (1984). "Equilibrium Unemployment as a Worker Discipline Device," American Economic Review 74 (June): 433-44

Michael Spence (1973). "Job Market Signaling," Quarterly Journal of Economics 88: 355-374.

Michael Spence (1974). Market Signaling Cambridge: Harvard University Press.

Robert Sproule (2000). "Student Evaluation of Teaching: A Methodological Critique of Conventional Practices," Education Policy Analysis Archives 8 (50): Nov. 2.

Robert Sproule (2002). "The Underdetermination of Instructor Performance by Data from the Student Evaluation of Teaching," Economics of Education Review 21: 287-294

R. J. Stapleton and G. Murkison (2001). "Optimizing the Fairness of Student Evaluations: A Study of Correlations Between Instructional Excellence, Study Production, Learning Production, and Expected Grades." Journal of Management Education 25 (3): 269-291.

J. E. Stone (1995). "Inflated Grades, Inflated Enrollment, and Inflated Budgets: An Analysis and Call for Review at the State Level," Education Policy Analysis Archives 3 (11).

M. Theall and J. Franklin (2001). "Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction," New Directions for Institutional Research 27(5): 45-56.

M. Thornton (2004). "Does Academic Publishing Pass the Real Market Test?" Public Choice 120 (1-2), May: 41-61.

Trout, Paul (1999). "Deconstructing an Evaluation Form." http://www.bus.lsu.edu/accounting/faculty/crumbley/deconstructing_evaluation.html

Trout, Paul (2000). "Low Marks for Top Teachers." Washington Post, March 13: A17.

Wall Street Journal (2003). "Low Marks for High Marks." Sept. 5 Editorial: W15.

Penelope J. Yunker and James A. Yunker (2003). "Are Student Evaluations of Teaching Valid? Evidence From an Analytical Business Course." Journal of Education for Business 78 (6), July/Aug.: 313-317.