Computers & Education 53 (2009) 429–435

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using a multilevel growth modeling approach

Adam C. Carle *

Department of Psychology, University of North Florida, 1 UNF Drive, Jacksonville, FL 32224, United States

ARTICLE INFO

Article history:
Received 19 November 2008
Received in revised form 24 February 2009
Accepted 2 March 2009

Keywords:
Post-secondary education
Pedagogical issues
Students' evaluations of teaching effectiveness
Multilevel growth modeling
Cross-cultural differences

ABSTRACT

Aims: Do college students' ratings of a professor's teaching effectiveness suggest that a professor's teaching improves with time? Does anything predict which instructors receive the highest ratings or improve the fastest? And, importantly, do the correlates of change differ across face-to-face and online courses?
Methods: I used data from 10,392 classes taught by 1120 instructors across three years and fit a taxonomy of multilevel growth models to examine whether students' ratings of teaching effectiveness (SETEs) changed across time, whether differences in average SETEs correlated with growth, and whether online vs. face-to-face, tenure, discipline, course level, sex, or minority status affected these estimates.
Results: SETEs remained relatively stable across time and teachers, although analyses uncovered a statistically significant, negative correlation between initial status and growth. Instructors starting with lower SETEs improved the fastest. These findings held across online and face-to-face instruction modes. However, in face-to-face classes, minority instructors received significantly lower average SETEs. This difference did not occur in online classes. No other predictors showed statistically significant effects. Finally, considerable SETE variance remained unexplained even when including the full predictor set in the model.
Discussion: These findings reveal that professors' SETEs can improve. Additionally, they indicate that patterns of change in teaching effectiveness do not differ generally across online and face-to-face instruction modes. However, the results showed that minority teachers in face-to-face but *not* online classes received lower evaluations than their majority counterparts. Additional research should seek to understand what leads to SETE differences across minority and majority groups in face-to-face classes but *not* online classes.

1. Introduction

Do college students' ratings of a professor's teaching effectiveness suggest that a professor's teaching improves with time? Does anything predict which instructors receive the highest ratings or improve the fastest? And, importantly, do the correlates of change differ across face-to-face and online courses? While a relatively large literature supports and addresses the reliability, validity, and use of students' ratings of teaching effectiveness (SETEs) based assessments (Abrami & d'Apollonia, 1991; Abrami, d'Apollonia, & Rosenfield, 1997; Marsh, 1987, 2007a; Marsh & Dunkin, 1992; Marsh & Roche, 1997), less research has examined predictors of teachers' success at the college level as measured by SETEs. Studies have tended to focus on cross-sectional mean differences and have not well addressed whether SETEs change (grow or deteriorate) across time generally and whether instructor or class-level differences predict differences in growth (Marsh, 2007b). *No* studies have examined instruction mode (face-to-face vs. online) differences, at least at the college level.

Given SETEs' influence on promotion and tenure; the evidence that men outnumber women at the higher ranks of the professoriate (AAUP, 2004; Benjamin, 1999; Nettles, Perna, & Bradburn, 2000); the observation that men and women differ in their distribution across the disciplines (Halpern, 2000); the data that white men generally earn more than white women, minority men, and minority women (AAUP, 2004; Benjamin, 1999; Nettles et al., 2000); and the increase in online courses (Moore & Anderson, 2003); it seems unconscionable not to examine these variables' relation to SETEs (Anderson & Smith, 2005). If students systematically rate minority professors more poorly than

* Tel.: +1 904 620 3573; fax: +1 904 620 3814.
E-mail address: adam.carle@unf.edu

0360-1315/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2009.03.001

white professors, women more poorly than men, one field more highly than another, or one instructional mode higher than another, this may partly explain salary differences. It would call into question SETEs' overall validity. It would also offer insight into how students and professors interact and lead to targeted efforts to decrease disparities in teaching effectiveness based on theoretical explanations for any observed differences (Anderson & Smith, 2005; Basow, 1987, 1995; Marsh, 1987, 2007a; Marsh & Roche, 1997; Smith & Anderson, 2005).

Some studies have investigated the relation of instructor- and class-level predictors to SETEs. However, these studies have primarily examined cross-sectional mean differences (e.g., mean differences at a single point in time), generally focused on one variable (e.g., sex), and leave a mixed impression for the role that class and instructor level predictors play in SETE prediction. For example, cross-sectional studies investigating whether an instructor's sex corresponded to differences in student evaluations initially suggested that women professors receive lower SETEs than men (Basow, 1987, 1995). But, Basow (1987, 1995) found small effect sizes and later studies have mostly failed to replicate differences (Anderson & Smith, 2005; Feldman, 1993; Ikegulu & Burham, 2001; Smith & Anderson, 2005). Moreover, studies that have replicated sex differences have reproduced small effect sizes that favor women, mean $r = 0.02$ (see Feldman (1993) for a review). Likewise, a similar pattern of findings has presented when examining seniority and course level (Blackburn & Lawrence, 1986; Feldman, 1977, 1983, 1997; Horner, Murray, & Rushton, 1989; Marsh, 1987, 2007a, 2007b; Marsh & Dunkin, 1997; Marsh & Hocevar, 1991; Murray, 1997; Renaud & Murray, 1996). Differences occasionally present, though not always, and the differences usually have little substantive meaning. And, although a relative paucity of studies has investigated the possible role a course's discipline may play in student ratings, similar findings result; some cross-discipline differences may exist, but the means don't usually differ meaningfully in size (Beran & Violato, 2005). Similarly, though my review found no studies directly addressing effectiveness across online and face-to-face classes, Lowerison, Sclater, Schmid, and Abrami (2006) found no differences in SETEs across classes that integrated computers and those that did not. On the other hand, those studies exploring minority/majority based differences have robustly found that minority instructors receive lower SETEs than the white (majority) counterparts and they routinely find larger effect sizes than those observed between men and women or across disciplines (Andersen & Miller, 1997; Anderson & Smith, 2005; Ludwig & Meacham, 1997).

These mainly cross-sectional studies suggest that course and instructor level variables may predict some mean SETE differences, but, excepting ethnicity, the size of the differences range close to zero and may reflect sample characteristics. Nearly all of these investigators called for replications and extension. None directly investigated differences across online and face-to-face classes. Moreover, few of these studies investigated change across time and, apart from one, none explicitly modeled the nested nature characteristic of educational data (Marsh, 2007b; Raudenbush & Bryk, 2002). Here, nested refers to the hierarchical nature of much education data. Datasets often include multiple observations of a single teacher at a single point in time (i.e., a single professor often teaches more than one class a semester) and to the fact that longitudinal datasets include multiple observations of a single teacher across time (i.e., a single professor generally teaches several classes across the observational time period). Thus, many observations are nested within a single professor and the data have a hierarchical structure.

These problems leave many questions unanswered. For example, cross-sectional analyses may show no sex, discipline or seniority differences, but longitudinal analyses might. Likewise, it remains unknown whether online and face-to-face evaluations show different relationships across predictors and time. Finally, because failure to model nesting leads to inflated Type I errors (Singer & Willett, 2003), previously observed differences may correspond to methodological artifacts, particularly given the small observed effects. Thus, the complex longitudinal relations among time, instruction mode, sex, minority status, discipline, and SETEs remain relatively unclear. The field needs longitudinal studies that investigate these relations using the best possible methodology.

Multilevel growth models more fully examine change and rise to the challenge presented above (Collins & Sayer, 2001; Duncan, Duncan, & Stryker, 2006; Raudenbush & Bryk, 2002; Singer & Willett, 2003). Multilevel growth models simultaneously model nested cross-sectional data (e.g., some professors teach more than one class a semester), and longitudinally nested data (e.g., multiple observations of a single professor across time). As a result, multilevel growth models can investigate possibilities rarely examined in SETE data (Marsh, 2007b) and arrive at more precise inferential decisions while doing it (Raudenbush & Bryk, 2002; Singer & Willett, 2003). Specifically, multilevel growth models allow researchers to investigate: cross-sectional mean SETEs at a study's start, growth (or decay) in SETEs across time and professors, the correlations between initial SETE status and growth, and differences for a professor within and across time. Importantly, they allow scientists to explore whether any variables predict differences in these relations (e.g., do face-to-face and online professors differ in their rates of growth). Only one study has used multilevel growth modeling in the SETE context. Marsh (2007b) found that: SETEs showed no improvement or deterioration with time (professors received similar SETEs across time), little variation existed for a given professor (professors received similar ratings across their classes at a given time point), initial status did not correlate with rates of change (professors starting the study with lower SETEs did not improve at a faster or slower rate than peers with higher SETEs), and instructors varied systematically in their SETEs (some professors consistently received higher SETEs than others). These results held across class level (graduate vs. undergraduate) and career statuses (early vs. late). Finally, Marsh (2007b) noted a relatively large proportion of variance remained unexplained and specifically called for a replication of these findings in new data. Though he did not address it, this replication could easily incorporate additional predictors in its multilevel analysis.

In sum, previous research has often used a cross-sectional approach that ignored the nested nature of SETE data and either aggregated the data (taken a teacher average across classes), which limits inferential conclusions, or proceeded as if the observations occurred independently, which leads to an increased Type I error likelihood (Singer & Willett, 2003). Given many consistently small effect sizes in the SETE literature, models that do not fully account for the nested nature of data may identify small, sample based deviations as reliable when the differences reflect only random sampling differences. Because differences across men and women, ethnicity, and discipline have broad policy and substantive implications; fresh efforts should replicate and extend multilevel growth model findings and examine whether class and instructor level predictors predict longitudinal differences.

In the current study, I addressed this set of concerns. I systematically replicated contemporary work suggesting little change in students' evaluations of teaching effectiveness (Marsh, 2007b) and extended this evaluation across online and face-to-face classes, as well as across several class and instructor level variables (instructor's sex, ethnicity, and tenure status, as well as the course's discipline and level). Using data from 10,392 classes (97% face-to-face, 3% online) taught by 1120 instructors across three years, I fit a taxonomy of multilevel growth models that examined college students' self-reported overall satisfaction with an instructor's teaching effectiveness at a course's end. First, for face-to-face classes, I predicted that I would replicate Marsh's (2007b) primary finding and observe little reliable growth among University professor's SETEs. Second, I also expected to find relatively substantial and systematic variance would exist across instructors. Some professors would consistently receive higher ratings than others. Third, I hypothesized that ethnicity would predict statistically significant differences in SETEs but that instructor's sex would not, nor would an instructor's tenure status or the course's level provide meaningful prediction. Given the paucity of findings relative to discipline or face-to-face vs. online teaching, I made no specific hypotheses.

2. Methods

2.1. Procedures

Multilevel growth models can use information from an individual even if they only appear once in a longitudinal data set (Singer & Willett, 2003). Thus, the data in this study come from the *entire* set of SETEs collected across seven semesters, three years (2005–2007), 10,392 classes, and 1120 instructors at a mid-sized, comprehensive, public, Southern United States University. The University collected data in a uniform manner. During each semester's final two class weeks (Fall, Spring, and Summer), students in all classes completed an institutionally standardized, mandatory Instructional Satisfaction Questionnaire (ISQ) that included 23 items. Using a polytomous scale (1 = poor and 5 = excellent), students rated several aspects of course satisfaction, including a single item that asked students to rate the professor's overall performance at the course's end. Professors distributed ISQs, read standardized instructions that included an assurance of anonymity, and left. Student volunteers collected ISQs, brought them to a central processing location, and the Office of Institutional Research (OIR) independently scored the results. Online classes proceeded similarly, except students anonymously provided ratings online. Upon processing the entire University's ISQs, the OIR provided professors the results and archived the data at the class level (class item means).

2.2. Measures

Student Evaluations of Teaching Effectiveness (SETEs): All analyses used a class' mean rating on a five point polytomous item (1 = poor and 5 = excellent) assessing the professor's overall performance at the course's end.

Instruction Mode: A single variable identified courses as online or face-to-face.

Course Level: A single variable coded courses as undergraduate or graduate.

Discipline: The OIR's archive included a course's subject area. To investigate possible discipline effects, I initially compared Science classes (broadly defined) to all other disciplines. Science included: Biology, Physics, Chemistry, all Mathematics, all Engineering, and all Social Sciences. All others received an Other code. Additionally, I compared: Social to Other Sciences, Physical to Life Sciences, Social Sciences to Humanities, Non-Social Sciences to Humanities, Non-Social Sciences to Education, and Social Sciences to Education. I also made several multi-group contrasts: Non-Social Science, Social Science, and Other; Life, Physical, and Social Sciences; Life Science, Physical Science, and Humanities; and Life Science, Physical Science, and Education. All of these contrasts revealed the same findings (reported below). Thus, for parsimony and to conserve space, I report only the initial Science to all other disciplines results. I will gladly share all results.

Tenure Status: Institutional records indicated professors' tenure status upon their initial inclusion in the study. I coded for tenure and not-yet tenured and excluded non-tenure track professors from analyses including this variable.

Minority Status: Institutional records indicated a professor's ethnicity as: White (81%), Asian (5%), Black (4%), Hispanic (4%), or Other (6%). I investigated: White (majority) compared to Hispanic, White compared to Black, White compared to Asian, and majority (White) compared to minority (all non-White). All of these contrasts revealed the same effect (reported below). Thus, for parsimony and to conserve space, I report only the majority compared to minority results. I will gladly share all findings.

Sex: University records designated a professor's sex (men = 58%, women = 42%).

2.3. Analytical approach

Given the multilevel nature of the SETE data, I followed methods described by Singer and Willett (2003) and fit a taxonomy of multilevel longitudinal growth models to investigate the questions of interest. These models included the facts that: the data came from multiple instructors measured several times and, at a given time point, most teachers taught more than one class and had multiple SETEs at each time point. Thus, the multilevel model allowed for dependence longitudinally *and* cross-sectionally and resulted in appropriately estimated standard errors and statistical significance tests (Singer & Willett, 2003). I conducted all analyses using full maximum likelihood (FML) estimation in SAS Proc Mixed (SAS Institute, 2001). I used an omnibus model-level alpha of 0.05. To control for inflated Type I error in post hoc tests, I used an alpha of 0.01. I coded the predictors to aid substantive interpretations (Singer & Willett, 2003) as follows: time (rectilinear[1]: Time = 0,…,6), curvilinear (quadratic) time (Time "squared"), instruction mode (0 = face-to-face, 1 = online), course level (0 = undergraduate, 1 = graduate), sex (0 = Male, 1 = Female), discipline (0 = Non-Science, 1 = Science), tenure status (0 = Tenured, 1 = Not Tenured), and minority status (0 = Majority, 1 = Minority)[2], and centered the criterion (SETE) measure. Finally, I examined the tenability of model assumptions, including error covariance assumptions, using the methods described by Singer and Willett (2003) and those available in SAS Proc Mixed (SAS Institute, 2001). None of the models demonstrated problematic violations and the "standard" error covariance assumptions proved most parsimonious and fit the data well (Singer & Willett, 2003).

3. Results

3.1. Multilevel analyses

I first fit a three-level unconditional means model (level-3 = teacher; level-2 = time; level-1 = class) that included no predictors in the full sample. However, analyses revealed a negative variance component for the class-level random effect suggested that little variation existed among classes within teachers (Singer & Willett, 2003). One can address this by simplifying the model and setting the variance

[1] Hereafter, I refer to rectilinear time as "time" or "linear time".

[2] As I discuss above and below, I examined discipline and minority status in a number of different manners. This reflects the final model.

component for the class-level random effect to zero. This approach fits a taxonomy of simpler two-level longitudinal growth models that does not model variation among classes within teachers (Singer & Willett, 2003). Alternatively, a negative covariance component can result from unbalanced data (Singer & Willett, 2003). For example, not all professors taught the same number of classes at each time. Thus, one should first empirically explore whether the negative covariance resulted from unbalanced data. If analyses in balanced data reveal significant variation among classes within teachers, this suggests that the negative covariance results from unbalanced data. In this case, one should fit the more complex model using the balanced data. However, if analyses in balanced data reveal little variation among classes within teachers, this suggests fitting the reduced two-level models (Singer & Willett, 2003).

I took this approach. I first created a balanced dataset and selected from the larger sample only those instructors with multiple measures at every data collection wave ($n = 446$). I then fit the unconditional three-level model in this subsample. The variance component describing residual variance among classes within teachers did not differ significantly from zero ($\hat{\sigma}^2_{\varepsilon\text{Class}} = 0.02$, SE = 0.05, $p < 0.33$). This suggested that the negative variance component for the class-level random effect in the full sample primarily resulted because most teachers received similar ratings across their classes. This supported setting the variance component for the class-level random effect to zero and fitting a taxonomy of reduced two-level longitudinal growth models. I did this. All subsequent models used the reduced two-level taxonomy.

Model 1: Unconditional Means Model: Model 1 included no predictors and supplied a baseline for future model comparisons (Singer & Willett, 2003). The variance components showed statistically significant variance associated with teachers ($\hat{\sigma}^2_0 = 0.219$, SE = 0.01, $p < 0.001$) and statistically significant residual variance ($\hat{\sigma}^2_s = 0.188$, SE = 0.003, $p < 0.001$). The intraclass correlation coefficient formed from the variance components showed that approximately 54% SETE variance arose from differences among teachers. Table 1 summarizes these results.

Model 2: Unconditional Growth Model: Model 2, the ‘‘standard” unconditional growth model (Singer & Willett, 2003), incorporated fixed and random effects for linear time. This resulted in statistically significant fit improvement ($\Delta^2_\chi(3) = 94.70$, $p < .01$), as well as reduced AIC and BIC. Model 2's variance components revealed statistically significant variance associated with teachers ($\hat{\sigma}^2_0 = 0.23$, SE = 0.01, $p < 0.01$) and statistically significant residual variance ($\hat{\sigma}^2_\varepsilon = 0.180$, SE = 0.003, $p < 0.01$). The linear fixed effect of time differed significantly and positively from zero, but slightly ($\gamma_{10} = 0.008$, SE = 0.002, $p < 0.01$). The variance component associated with time indicated small, statistically significant, systematic differences among linear trends across instructors ($\hat{\sigma}^2_1 = 0.002$, SE < 0.001, $p < 0.001$), and the residual covariance component differed significantly from zero ($\hat{\sigma}^2_{01} = -0.006$, SE = 0.002, $p < 0.01$). Expressed as a correlation coefficient, $\hat{\rho}_{\pi_0\pi_1} = \sqrt{(\hat{\sigma}^2_0\hat{\sigma}^2_1)} = -0.27$, it suggested a small to moderate (Cohen, 1988) negative correlation between initial status and rates of change. Teachers with initially low intercepts improved relatively quickly compared to individuals with higher intercepts. Finally, adding the fixed and random effects for time accounted for 4% of the explainable variance.

Model 3: Quadratic Effect of Time: Before exploring additional predictors, I examined the possibility that change across time did not follow a rectilinear (straight line) form. Results did not support this contention. The estimate associated with the fixed effect of quadratic time did not differ significantly from zero, the addition of the quadratic effects did not significantly reduce the deviance statistic ($\Delta\chi^2(4) = 4.6$, $p = 0.20$), and the AIC and BIC increased. Thus, I dropped this term. No remaining analyses included quadratic effects.

Model 4: Effect of Instruction Mode: Model 4 examined overall differences in initial status across instruction mode by adding a fixed effect for instruction to Model 2. Model 4 explored the possibility that online and face-to-face classes differed in their overall rates of change. It

Table 1

Results of selected growth models examining professors' SETEs across seven semesters, instruction mode (face-to-face and online), and minority status.[a]

	Parameter	Unconditional means	Unconditional growth fixed and random	Reduced model
Initial status				
Intercept	γ_{00}	−0.025	−0.051*	−0.007
(SE)		(0.015)	(0.018)	(0.019)
Mode	γ_{01}	–	–	0.032
(SE)		–	–	(0.138)
Ethnicity	γ_{02}	–	–	−0.266*
(SE)		–	–	(0.048)
Mode ∗ ethnicity	γ_{03}	–	–	−0.264
(SE)		–	–	(0.294)
Rate of change				
Intercept	γ_{10}	–	0.008	0.008*
(SE)		–	0.002	(0.003)
Mode	γ_{11}	–	–	−0.033
(SE)		–	–	(0.030)
Ethnicity	γ_{12}	–	–	0.010
(SE)		–	–	(0.007)
Mode ∗ ethnicity ∗ time	γ_{112}	–	–	−0.017
(SE)		–	–	(0.064)
Level 1 (residual)				
Within person	σ^2_ε	0.188*	0.180*	0.180*
(SE)	Residual	0.003	(0.003)	(0.003)
Level 2				
Initial status	σ^2_0	0.219*	0.233*	0.224*
(SE)		0.011	(0.014)	(0.013)
Rate of change	σ^2_1	–	0.002*	0.002*
(SE)		–	(0.000)	(0.000)
Rate of change covariance	σ^2_{01}	–	−0.006*	−0.005*
(SE)		–	(0.002)	(0.002)

* $p < 0.01$.

[a] Mode reference group = face-to-face. Ethnicity reference group = majority.

also examined whether covariation remained between initial status and rates of change, controlling for possible mode differences by adding the fixed effect interaction between mode and time. Model 4 failed to find any statistically significant fixed effect mode differences indicating that professors teaching face-to-face and online classes received similar average SETEs and that SETEs had a similar relationship with time across mode of instruction. I retained this variable in the model because the next series of models explored whether sex, discipline, tenure, or minority status predicted differences across instruction mode.

Models 5 and 6: Effect of Sex, Discipline, Tenure, and Minority Status across Instruction Mode: In Model 5, I explored the possibility that sex, discipline, tenure, or minority status predicted differences in SETEs. To do this, Model 5 added fixed effects for these variables to Model 4. This examined differences in initial status across predictors controlling for instruction mode. I also explored the possibility that sex, discipline, tenure, or minority status predicted different rates of change and whether covariation remained between initial status and rates of change by adding the fixed effect interactions between each predictor (sex, discipline, tenure, and minority status) and time to Model 4. Finally, to examine whether any of these relationships differed across instruction mode, I included interactions between each predictor and instruction mode.

Model 5 exhibited statistically significant residual and teacher variance components, $\hat{\sigma}^2_\varepsilon = 0.18$ (SE = 0.004, $p < 0.01$) and $\hat{\sigma}^2_0 = 0.16$ (SE = 0.02, $p < 0.01$), respectively. The linear fixed effect of time displayed small systematic differences among instructors ($\hat{\sigma}^2_1 = 0.002$, SE < 0.001, $p < 0.01$). Considering the variables in the model, the residual covariance suggested a small correlation between initial status and rates of change. With respect to predictors, Model 5 revealed a moderate effect for minority status ($\hat{\gamma}_{01} = -0.24$, SE = 0.06, $p < 0.01$) controlling for other variables. On average in face-to-face classes, minority instructors received lower SETEs than their majority peers. Results indicated that majority and minority professors' SETEs *did not* differ significantly in online classes. The majority/minority difference occurred *only* in face-to-face classes. No other predictors differed significantly.

To encompassingly investigate minority status, I also considered models with different racial and ethnic comparisons: White vs. Hispanic, White vs. Black, White vs. Asian, and White vs. Other, all resulted in similarly patterned and sized findings. I adopted the majority/minority comparison as the most parsimonious and present only those results. I also did this for using several different discipline comparisons (see above). Each resulted in similarly patterned and sized null findings for discipline. I will share all analyses' details upon request.

For interpretive parsimony (Singer & Willett, 2003), I estimated a final model (Model 6) that included only ethnicity and instruction mode. Model 6 showed statistically significant residual ($\hat{\sigma}^2_\varepsilon = 0.18$, SE = 0.003, $p < 0.01$) and teacher variance components ($\hat{\sigma}^2_0 = 0.224$, SE = 0.013, $p < 0.01$). Its linear fixed effect of time displayed small systematic differences among instructors ($\hat{\sigma}^2_1 = 0.002$, SE < 0.001, $p < 0.01$). Considering the variables in the model, the residual covariance suggested a small correlation between initial status and rates of change ($\hat{\rho}_{\pi_0\pi_1} = -0.26$). Adding the effect of minority status accounted for 4% of the explainable variance and 6% of variance among intercepts, it resulted in statistically significantly improved model fit statistic ($\Delta\chi^2(3) < 43.9$, $p = 0.01$), and decreased the AIC and BIC (see Table 1). Like Model 5, Model 6 indicated a moderate effect for minority status ($\hat{\gamma}_{01} = -0.27$, SE = 0.048, $p < 0.01$) controlling for other variables in the model. Face-to-face minority instructors received lower SETEs than their majority peers on average ($d = -0.43$). This effect did not occur for online classes. Table 1 summarizes these results.

4. Discussion

What do college students' ratings of teaching effectiveness (SETEs) across time and instruction mode (online vs. face-to-face) reveal? In an effort to address this question, I used a multilevel growth modeling approach and uncovered several intriguing findings. First, like others (Marsh, 2007b), I found that SETEs showed little improvement across time. Professors who received high ratings at the study's inception received similar ratings at the study's conclusion. Those who started low tended to end with similarly depressed SETEs. Though I did observe some positive and reliable growth (instructors did tend to receive higher SETEs as they gained experience), it occurred minimally. For each additional semester of teaching, SETEs improved an average of 0.01 points on a five point scale. Thus, across 25 years (including summers), the average professor improves approximately 0.6 points on the scale. Professors with SETE's well below average (2 STD) improve approximately 1 point. Perhaps more relevant for tenure decisions, across a five year time period the average professor improves 0.12 points and professor's well below average improve 0.5 points.

Relatedly, I found no evidence for curvilinear change. The little change that occurred followed a linear model. Moreover, little variance presented in SETEs across an instructor's classes at a given point in time. Although professors taught diverse classes across different levels, individual instructors tended to receive similar ratings across their classes. Interestingly, I did observe a correlation between initial SETE status and growth. Those who started the study with lower ratings tended to improve most rapidly ($r = -0.26$). This suggested that instructors receiving relatively poor SETEs adjusted their teaching more quickly than those who received higher ratings. However, little change occurred generally. Finally, these findings held across face-to-face and online classes. Changes in professors' teaching effectiveness in face-to-face and online classes appear to follow similar patterns.

Yet, to extend previous work, I pursued several possible predictors of systematic variance across teachers' SETEs indicating that some instructors consistently received better ratings than others. I observed *no* effects for sex, discipline, tenure status, or course level. At the study's inception, men and women's SETEs did not differ significantly and their SETEs showed similar stability. Likewise, across disciplines, course level, and tenure status, SETEs generally evidenced stable and similar ratings. This held across face-to-face and online classes. In both face-to-face and online classes it appears that sex, discipline, tenure status, and course level fail to predict students' ratings of a professor's effectiveness.

This may leave one wondering, did anything predict differences among professors? Yes. In face-to-face classes, minority status proved a statistically significant predictor of mean SETE difference and showed a relatively moderately sized effect. Minority professors received SETEs −0.4 points lower than their majority (White) peers on average ($d = -0.43$). Relatedly, across time, minority and majority instructors' SETEs displayed similar rates of change (i.e., both groups had little growth), assuring the differences between these groups persisted across the study. This finding held across numerous majority/minority-based comparisons (e.g., White compared to Hispanic, White compared to Black, White compared to Asians, etc.). It held when including course level, tenure status, and discipline in the model. And, when I included ethnicity in the model, the other predictors continued to contribute little. However this finding did *not* hold across face-to-face and online classes. In online classes, SETEs did not differ across majority and minority professors. In online classes, a professors' ethnic status did *not* influence students' rating of a professor's teaching effectiveness.

What might account for these findings? Previous explanations of similarities and differences in SETEs have generally rested on sexism (Basow, 1987, 1995; Feldman, 1993) or racism (Anderson & Smith, 2005). They primarily argued that when professors violate social norms students perceive and subsequently rate them poorly. However, because the examination here represents one of the first to investigate the influence of several variables simultaneously within the multilevel framework, the explanation needs to concomitantly address both null and difference findings. Thus, while previous research has investigated these variables, many of these studies investigated sex or ethnicity in isolation and accounts rarely concurrently addressed null sex findings and statistically significant differences across ethnicity. These studies did not address the fuller and more precise set of empirical inferences allowed in longitudinal, multilevel models. And, *none* examined the presence or lack of an effect across instruction mode. Consequently, previous discussions lack broader explanative power and the findings may spuriously reflect the populations of interest. A satisfactory explanation of the current findings must simultaneously address why minority instructors received lower ratings but women did not and it must address differences across face-to-face and online courses. The field of perceived similarity provides a possible explanation.

The domain of perceived similarity and self-effectiveness has shown that individuals generally trust, appreciate, aid, and listen to others they regard similarly to themselves (Bandura, 1993; Moss, Garivaldis, & Toukhsati, 2007). The theory suggests that individuals who perceive themselves as similar will simultaneously like and support each other. When individuals perceive themselves as dissimilar (e.g., majority students interacting with minority professors), the opposite effect will occur (Liao, Joshi, & Chuang, 2004). Moreover, when individuals (students) perceive themselves as similar to a model (professor), the model functions more efficaciously (Bandura, 1977, 1989, 1993, 1995). Thus, when students perceive instructors as more like themselves they may generally rate the instructor more highly, as well as like and support the professor more. Increased support from students should lead to more effective teaching. Likewise, teachers who view themselves efficaciously teach better (Ashton & Webb, 1986; Bandura, 1993; Chwalisz, Altmaier, & Russell, 1992; Gibson & Dembo, 1984). When professors perceive students as similar to themselves, they may feel and subsequently teach more effectively (Bandura, 1993). Finally, when students perceive an instructor as similar to themselves, the learning that occurs as a result of modeling should occur more effectually. In sum, perceived similarities should generally lead to higher SETEs. Perceived dissimilarities should correspond to the opposite.

This fits with the similarities and differences observed here. Despite heterogeneity, the campus consisted mostly of white students (75%) and professors (81%). Accordingly, minority professors generally taught majority students who, in turn, may have perceived minority instructors as dissimilar to themselves. Likewise, minority professors teaching mostly majority students may have perceived their students as dissimilar. Both could lead to reduced SETEs. On the other hand, the university had a relative balance of the sexes across students (58% female) and faculty (42% female) resulting in ample opportunities for perceived similarities and the subsequent benefits across sex. Finally, in the online environment, which likely removes obvious physical difference cues, the effect of minority status disappeared. Minority professors no longer received lower evaluations. Removing a perceived dissimilarity cue corresponded to more equal ratings. Consistent with an extensive literature describing the advantages of online teaching to teachers and students (Jaffee, 2003; Moore & Anderson, 2003, 2007), this indicates the possibility that online classes may present a more egalitarian environment than face-to-face classes.

Before concluding, the study's limits and strengths merit review. First, the study included no student specific predictors. Because the University archived at the class level, I did not have individual student information and could not investigate student-level differences. However, Marsh (2007a, 2007b; Marsh & Roche, 1997) has strongly argued that class-level aggregations generally provide more valid indicators of teaching effectiveness than student-level SETEs regardless. Second, the archived data allowed no class-level predictors beyond instruction mode, course level, and discipline, nor did additional data exist to define various forms of online teaching. Additional class-level predictors might explain some of the observed heterogeneity. Finally, although the data replicate Marsh's (2007b) work, this pattern of results may not hold in all samples. One should take some care when generalizing these results to all Universities.

Despite these limits, the study has numerous assets. First, the large sample size and inferential precision offered by the multilevel growth modeling approach allowed me to powerfully and uniquely examine a number of new questions (e.g., did initial status and rate of change relate). Second, this study extended previous findings to a new sample, helping to establish generalizability. Third, the study broadened previous work by uniquely examining the longitudinal relation of class- and teacher-level variables simultaneously to SETEs across face-to-face and online classes. Because, teaching effectiveness interventions assume that professors can improve, the fact that I observed systematic growth in both face-to-face and online classes suggests that professors can and do improve with time.

5. Conclusion

In summary, the current study used multilevel growth modeling and a large heterogeneous sample of college professors at a comprehensive, public University to investigate student perceptions of teaching effectiveness across time and face-to-face and online courses. Across seven semesters' evaluations, I found little change in SETEs among instructors teaching face-to-face or online courses. The growth that did arise occurred minimally and held robustly across instruction mode, discipline, sex, course level, tenure status, and ethnicity. However, in face-to-face classes *only*, students consistently and stably gave lower SETEs to minority professors. They did not differentially rate professors as a function of the professor's sex, tenure status, discipline, or level of the course the professor taught. Students did *not* rate minority professors more poorly in online courses. These findings indicate that online instructors do not receive a penalty or boost relative to professors teaching face-to-face classes. Additionally, the fact the minority professors did not receive poorer ratings suggests that online classes may function more impartially. Summarily, the present study contributes to the field's understanding of differences and similarities across SETEs. Future studies should seek to understand what causes differences across minority status in face-to-face and online courses and it should aim to elucidate the remaining unexplained variance in SETEs that occurred regardless of instruction mode.

Acknowledgements

I would like to thank Tara J. Carle and Margaret Carle whose unending support and thoughtful comments make this work possible. I would also like to thank David Jaffee for his insightful comments on an early draft of this manuscript. Finally, I would like to thank Steven R. Howe, a truly inspiring and effective professor and teacher.

References

Abrami, E. C., & d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness – Generalizability of "$N = 1$" research: Comment on Marsh (1991). *Journal of Educational Psychology, 30*, 221–227.

Abrami, E. C., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: What we know and what we do not. In R. E. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321–367). New York: Agathon Press.

American Association of University Professors (2004). *Don't blame faculty for high tuition: The annual report on the economic status of the profession 2003–2004*. <http://www.aaup.org/AAUP/comm/rep/Z/ecstatreport2003-04/> Retrieved 14.07.08.

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *PS Political Science & Politics, 30*, 216–219.

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences, 184*, 201.

Ashton, P., & Webb, R. B. (1986). *Making a difference. Teachers' sense of efficacy and student achievement*. NJ: Longman.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 191*, 215.

Bandura, A. (1989). Human agency in social cognitive theory. *American Psychologist, 44*, 1175–1184.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117–148.

Bandura, A. (1995). *Self-efficacy in changing societies*. Cambridge University Press.

Basow, S. A. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308–314.

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*, 656–665.

Benjamin, E. (1999). Disparities in the salaries and appointments of academic women and men. *Academe, 85*, 60.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*, 593–601.

Blackburn, R. T., & Lawrence, J. H. (1986). Aging and the quality of faculty job-performance. *Review of Educational Research, 56*(3), 265–290.

Chwalisz, K., Altmaier, E. M., & Russell, D. W. (1992). Causal attributions, self-efficacy cognitions, and coping with stress. *Journal of Social and Clinical Psychology, 11*, 377–400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Collins, L. M., & Sayer, A. G. (Eds.). (2001). *New methods for the analysis of change*. Washington, DC: American Psychological Association.

Duncan, T. E., Duncan, S. C., & Stryker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). London: Erlbaum.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 6*, 223–274.

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. *Research in Higher Education, 18*, 3–124.

Feldman, K. A. (1993). College students views of male and female college teachers: Part II evidence from students evaluations of their classroom teachers. *Research in Higher Education, 34*, 151–211.

Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon Press.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*, 569–582.

Halpern, D. F. (2000). *Sex differences in cognitive abilities*. Mahwah, NJ: L. Erlbaum Associates.

Horner, K. L., Murray, H. G., & Rushton, J. P. (1989). Relation between aging and rated teaching effectiveness of academic psychologists. *Psychology and Aging, 4*, 226–229.

Ikegulu, N. T., & Burham, W. A. (2001). Gender roles, final course grades, and faculty evaluation. *Research and Teaching in Developmental Education, 17*, 53–65.

Jaffee, D. (2003). Virtual transformation: Web-based technology and pedagogical change. *Teaching Sociology, 31*(2), 227–236.

Liao, H., Joshi, A., & Chuang, A. (2004). Sticking out like a sore thumb: Employee dissimilarity and deviance at work. *Personnel Psychology, 57*, 969–1000.

Lowerison, G., Sclater, J., Schmid, R. F., & Abrami, P. (2006). Student perceived effectiveness of computer technology use in post-secondary classrooms. *Computers & Education, 47*(4), 465–489.

Ludwig, J. M., & Meacham, J. A. (1997). controversial courses: Student evaluations of instructors and content. *Educational Research Quarterly, 21*, 27–38.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W. (2007a). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–384). New York: Springer.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99*, 775–790.

Marsh, H. W., & Dunkin, M. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.). *Higher education: Handbook on theory and research* (Vol. 8, pp. 143–234). New York: Agathon Press.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9–18.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*, 1187–1197.

Moore, M. G., & Anderson, W. G. (2003). *Handbook of distance education*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Moore, M. G., & Anderson, W. G. (2007). *Handbook of distance education*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Moss, S. A., Garivaldis, F. J., & Toukhsati, S. R. (2007). The perceived similarity of other individuals: The contaminating effects of familiarity and neuroticism. *Personality and Individual Differences, 43*, 401–412.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *International Journal for Academic Development, 2*(1), 8–23.

Nettles, M. T., Perna, L. W., & Bradburn, E. M. (2000). Salary, promotion, and tenure status of minority and women faculty in U.S. colleges and universities. *Education Statistics Quarterly, 2*, 94–96.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Renaud, R. D., & Murray, H. G. (1996). Aging, personality, and teaching effectiveness in academic psychologists. *Research in Higher Education, 37*, 323–340.

SAS Institute (2001). *Statistical analysis system*. <http://www.sas.com>.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. NY, NY: Oxford University Press.

Smith, G., & Anderson, K. J. (2005). Students' ratings of professors: The teaching style contingency for Latino/a professors. *Journal of Latinos and Education, 4*, 115–136.