2014

Forecasting the Student–Professor Matches that Result in Unusually Effective Teaching

Jennifer Gross
Grand Valley State University, grossj@gvsu.edu

Brian Lakey
Grand Valley State University

Jessica L. Lucas

Ryan LaCross

Andrea R. Plotkowski

See next page for additional authors

British Journal of Educational Psychology (2014)

The British Psychological Society

www.wileyonlinelibrary.com

Forecasting the student–professor matches that result in unusually effective teaching

Jennifer Gross*, Brian Lakey, Jessica L. Lucas, Ryan LaCross, Andrea R. Plotkowski and Bo Winegard

Grand Valley State University, Allendale, Michigan, USA

Background. Two important influences on students' evaluations of teaching are relationship and professor effects. Relationship effects reflect unique matches between students and professors such that some professors are unusually effective for some students, but not for others. Professor effects reflect inter-rater agreement that some professors are more effective than others, on average across students.

Aims. We attempted to forecast students' evaluations of live lectures from brief, video-recorded teaching trailers.

Sample. Participants were 145 college students (74% female) enrolled in introductory psychology courses at a public university in the Great Lakes region of the United States.

Methods. Students viewed trailers early in the semester and attended live lectures months later. Because subgroups of students viewed the same professors, statistical analyses could isolate professor and relationship effects.

Results. Evaluations were influenced strongly by relationship and professor effects, and students' evaluations of live lectures could be forecasted from students' evaluations of teaching trailers. That is, we could forecast the individual students who would respond unusually well to a specific professor (relationship effects). We could also forecast which professors elicited better evaluations in live lectures, on average across students (professor effects). Professors who elicited unusually good evaluations in some students also elicited better memory for lectures in those students.

Conclusions. It appears possible to forecast relationship and professor effects on teaching evaluations by presenting brief teaching trailers to students. Thus, it might be possible to develop online recommender systems to help match students and professors so that unusually effective teaching emerges.

Nearly all colleges and universities in the United States use students' evaluations of teaching as part of tenure and promotion decisions. Many measures of students' evaluations have impressive validity. For example, professors' scores on students' evaluations correlate substantially with students' learning, at least when assessed in courses with standardized examinations and content (Cohen, 1981; Feldman, 1989a; Marsh, 1984, 2007). Furthermore, there is a reasonable agreement between current students, faculty, administrators, and alumni about which professors are most effective

**Correspondence should be addressed to Jennifer Gross, Department of Psychology, Grand Valley State University, Allendale, MI 49401, USA (email: grossj@gvsu.edu).*

DOI:10.1111/bjep.12049

(Centra, 1975; Feldman, 1989b; Marsh, 1984, 2007). However, teaching evaluations have not yet been used to help match students and professors so that unusually effective teaching emerges. The goal of the current research was to test whether brief video trailers of professors' teaching can forecast students' evaluations of live lectures months later. We envision an online system similar in some ways to recommender systems used by Amazon.com and itunes that make individualized recommendations for music or book purchases. As applied to college teaching, a student would view and rate brief videos of professors' teaching and would then be given individualized feedback about which of these professors the student would find especially effective, based on the students' ratings as well as the ratings of other students. We believe such forecasting would be useful, regardless of whether the reader is persuaded of the construct validity of students' evaluations of teaching. If teaching evaluations reflect student learning, then forecasting evaluations should improve student learning by helping each student choose professors who are uniquely effective for the student. If teaching evaluations reflect only consumer satisfaction, then forecasting evaluations should lead to more satisfied students.

When a student rates a professor's effectiveness, the rating reflects at least three distinct influences (Gross, Lakey, Edinger, Orehek, & Heffron, 2009). Part of the student's rating reflects the objective effectiveness of the professor, as reflected in inter-rater agreement among observers that some professors are more effective than others (professor effects). For example, raters might consistently give Professor Hendersen higher ratings than Professor Duren. Professors' effectiveness (as measured by student-rater agreement) is the most widely studied aspect of student evaluations of teaching, and there is a good agreement about effectiveness across observers (Centra, 1975; Feldman, 1989b; Marsh, 1984, 2007). A second determinant of a student's rating is rater bias and occurs when some students characteristically rate the same professors more favourably than do other students, regardless of the actual characteristics of professors. For example, when Ellen and Will rate the same professors, Ellen gives higher scores than does Will, on average. A third determinant of a student's ratings are relationship effects (Kenny, 1994). Relationship effects occur when a student rates a professor (1) more favourably than the student typically rates professors (rater bias) and (2) more favourably than the professor is typically rated by other students (professor effects). That is, in an ANOVA model, relationship effects are rater $\times$ professor interactions. For example, Ellen might rate Professor Hendersen more favourably than one would expect, given (1) Ellen's tendency to rate professors leniently (rater bias) and (2) Professor Hendersen's tendency to elicit favourable ratings from students on average. Thus, relationship effects reflect the extent to which a professor is unusually effective for a specific student.

There is a good reason to expect relationship effects in college teaching as research using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and the social relations model (SRM; Kenny, 1994; Kenny, Kashy, & Cook, 2006) has identified relationship effects in a wide range of human judgments, including leadership (Livi, Kenny, Albright, & Pierro, 2008), personality (Park, Kraus, & Ryan, 1997), social support (Lakey & Orehek, 2011), physician's cultural competency (Lucas, Lakey, Arnetz, & Arnetz, 2010), psychotherapist qualities (Lakey, Cohen, & Neely, 2008), family negativity (Cook, Kenny, & Goldstein, 1991), and parent–child attachment (Cook, 2000).

Large relationship effects have also been observed in students' evaluations of college teaching and quiz performance (Gross *et al.*, 2009). In Study 1, undergraduate and graduate students rated their professors after completing full semester courses. In Study 2, a new sample of students rated live lectures. In Study 3, another group of students rated video-recorded lectures. Large relationship effects on teaching evaluations were found in

all studies. On average (median), relationship effects accounted for 52% of the variance in students' evaluations of professors' teaching. Moreover, Studies 2 and 3 included quizzes based on each lecture, and relationship effects on teaching effectiveness were linked to relationship effects on memory for lecture. That is, when a student rated a lecture as unusually effective, he or she also performed unusually well on the quiz. Gross *et al.* (2009) also estimated professor and rater effects. Consistent with the research on inter-rater agreement on teaching evaluations (Centra, 1975; Feldman, 1989b; Marsh, 1984, 2007), professor effects accounted for 30% of the variance. Rater bias accounted for 17%. In short, personal tastes (relationship effects) account for the largest variance in students' ratings and forecast learning in the classroom.

Forecasting which student will find which professor unusually effective requires methods that can isolate relationship effects from professor and rater effects, and predictive accuracy must be established separately for each. First, we describe the simpler cases of predicting perceived teaching effectiveness from professor and rater effects. To predict how students on average would respond to a specific professor, one could base prediction on professor effects. If the results of teaching evaluations were available for every professor, students should select the professor with the highest evaluations. If teaching evaluations derived from real courses were not publically available, prediction could be based on students' averaged ratings of 'teaching trailers'. To predict how a given student would evaluate all professors on average (rater bias), one could take each student's average rating across all teaching trailers. Predicting rater bias is not likely to be useful in practice. Yet, as described in the discussion, understanding rater bias could be useful when using teaching evaluations for personnel decisions. For example, it would be helpful to know whether some majors were more generous raters than other majors. To forecast how the student Ellen would uniquely respond to Professor Hendersen (relationship effects), one would take Ellen's reaction to Hendersen's teaching trailer, with Hendersen's average score across all students (cf. professor effects) and Ellen's average rating across all trailers (cf. rater effects) removed.

Research on other constructs suggests that it should be possible to forecast relational teaching evaluations. For example, Veenstra *et al.* (2011) attempted to forecast relational perceived support on the basis of brief conversations between support providers and recipients, as well as brief video interviews with each provider. First, each recipient viewed a video recording of an interview with each provider. Next, each recipient had a brief conversation with each provider. Recipients and providers continued to meet for several weeks (Study 1) or several months (Study 2). Veenstra *et al.* (2011) could forecast which provider a recipient would ultimately see as unusually supportive on the basis of a brief conversation at $r \approx .45$, but not from the video interview. Similarly, Park *et al.* (1997) found that relationship effects on personality were somewhat stable over time, suggesting it should be possible to forecast them. Still, predictive accuracy for relationship effects might not be as strong as for professor and rater effects. Kenny's (2004) quantitative theory of person perception predicts that professor (i.e., target) effects should be predicted with very high accuracy ($r \approx .90$). Similarly, rater effects should be predictable with excellent accuracy as Veenstra *et al.* (2011) forecasted rater effects in perceived social support with high accuracy ($r \approx .90$).

The goal of this study was to explore the viability of using video trailers of professors' teaching to forecast which students would respond unusually well to specific professors' teaching. Early in the semester, students viewed brief video trailers depicting professors' teaching. Students rated the effectiveness of each professor's teaching in the trailer, and students rated their own affect during the trailer. We included affect because students'

affect has been strongly linked to teaching ratings (Fortunato & Mincy, 2003; Gross *et al.*, 2009). These measures of perceived teaching effectiveness and experienced affect were used to forecast students' evaluations of live lectures later in the semester. Consistent with Gross *et al.* (2009), we hypothesized (1) significant relationship, rater, and professor effects on student evaluations of teaching and on students' affect experienced when watching the video trailers and live lectures and (2) significant correlations between relational teaching evaluations and relational memory for lectures. Consistent with Veenstra *et al.* (2011), we hypothesized that (3) relationship effects on evaluations and experienced affect in response to the trailers would forecast relationship effects on teaching evaluations during the live lecture.

Method

Participants

One hundred and forty-five college students (74% female; mean age = 19) from three sections of introductory psychology taught by JG completed all measures during the course of a semester. Isolating relationship effects requires a design in which students rate the same professors (Kenny, 1994; Kenny *et al.*, 2006). Gathering complete data from students required their participation during seven classes in each of sections 1 (N = 47) and 2 (N = 50) and 10 classes in section 3 (N = 48). Still, 78% of students enrolled participated in all sessions. Students missed classes for a variety of reasons including a required retreat for an academic programme, illness, weather-related travel difficulties, and off-campus games for athletes. Our analytic procedures are intolerant to missing data, and so, students with incomplete data were excluded from the analyses.

Guest lectures and teaching trailers were obtained from 10, tenure-track professors (80% male; mean age = 44; range = 33–64) from a medium-sized state university in the Great Lakes region of the United States. All professors were native English speakers. Three professors taught in the first section, three in the second section, and four in the third section. Students viewed the trailers only for the professors who gave guest lectures in the students' section. One professor taught in two sections. We retained the professor in both the sections to boost statistical power for professor effects, after determining that this professor's appearance in both the sections did not influence the results.

Procedure

Students were shown 6-min video trailers of each professor. Students rated their own affect and the effectiveness of each professor's teaching in response to each trailer. Later in the semester (median = 8 weeks, range = 3–12), students heard a 40-min live lecture by each professor. Students again evaluated teaching and rated their own affect. Students completed a quiz on each lecture during the next class period.

Teaching trailers

In the year preceding the study, each professor was video-recorded teaching a representative 50-min class. The authors independently viewed each recorded lecture and identified the key features of each professor's style (e.g., sarcastic humour, enthusiasm, quick pace, eye contact, confidence). In subsequent meetings, a consensus description of each professor's style was developed. Finally, video passages were identified that reflected the consensus description and were compiled into the final 6-min trailers.

Measures

Teaching evaluations

Students completed the widely used Students' Evaluations of Educational Quality (Marsh, 1982), modified for use with teaching trailers (18 items) and live lectures (24 items). For example, the item 'You found the course intellectually challenging and stimulating' was modified to read 'You found the lecture intellectually challenging and stimulating.' For all measures, we calculated internal consistency separately for each of the effects of interest.[1] For the trailers, internal consistency reliability was .97 for rater, .99 for professor, and .93 for relationship effects. For live lectures, reliability was .98 for rater, .99 for professor, and .95 for relationship effects.

Affect

After each trailer and lecture, students completed the 20-item, Positive and Negative Affectivity Schedule (Watson, Clark, & Tellegen, 1988), perhaps the most widely used measure of affect in psychology. The measure includes two mostly independent subscales of positive and negative affect. Example items include 'interested' and 'excited' for positive affect and 'distressed' and 'nervous' for negative affect. For positive affect, reliability was .96 for rater, .99 for professor, and .89 for relationship effects (trailers), and .98 for rater, .99 for professor, and .90 for relationship effects (live lectures). For negative affect, reliability was .96 for rater, .62 for professor, and .90 for relationship effects (trailers), and .95 for rater, .93 for professor, and .79 for relationship effects (live lectures).

Quizzes

Students' memory for live lectures was assessed by 12-item multiple-choice quizzes administered during the next class period. Reliability was .91 for rater, .70 for professor, and .34 for relationship effects. Although there was much more random error in the relationship effect than desired, as described momentarily, the effect was still able to replicate the link between teaching evaluations and quiz performance reported by Gross *et al.* (2009).

Statistical analyses

First, we determined the extent to which there were rater, professor, and relationship effects for each of the study constructs. We analysed the data as a students × professors (nested within sections) × item design in VARCOMP within SPSS. Each factor was random. Students and professors were nested within sections, and each factor was crossed with items. Each section formed a level of the sections factor, each student formed a level of the raters factor, each professor formed a level of the professors factor, and each item aggregate formed a level of the items factor. We constructed two aggregates of odd

[1] *Internal consistency reliability was estimated using the following formulas derived from generalizability theory (Cronbach* et al., *1972) for which* r = *rater,* p = *professor,* i = *item,* r × p = *rater by professor (i.e., relationship effects),* r × i = *rater by item,* p × i = *professor by item,* r × p × i = *rater* × *professor* × *item, and* n_i = *number of items:* $\alpha_r = \sigma_r^2/(\sigma_r^2 + (\sigma_{r\times i}^2/n_i))$; $\alpha_p = \sigma_p^2/(\sigma_p^2 + (\sigma_{p\times i}^2/n_i))$ *and* $\alpha_{r\times p} = \sigma_{r\times p}^2/(\sigma_{r\times p}^2 + (\sigma_{r\times p\times i}^2/n_i))$. *These generalizability coefficients are interpreted in the conventional manner. For example,* α_r *is essentially Cronbach's alpha (Cronbach* et al., *1972) and indicates the expected squared correlation between participants' scores on the items administered and participants' scores on all possible, similar items.* α_r *is interpreted similarly except that the unit of observation is each student–professor dyad.*

and even items for each construct to reduce measurement error and the size of the design (Gross *et al.*, 2009). We analysed the study as a nested design to obtain better statistical power for professor effects. Analysing each section separately has the disadvantage of only 3 or 4 professors per study, whereas the nested design had 10 professors. We also analysed each section's data separately, and the results were very similar to those of the nested design. Analyses of each section's data are available by request. The design yielded nine effects: raters nested within section (i.e., raters:sections), professors:sections, items, sections, raters:sections × items, professors:sections × items, items × sections, raters:sections × professors:sections, and raters:sections × professors:sections × items. Relationship effects are reflected in the raters:sections × professors:sections effect. Effects involving items are typically viewed as measurement error and thus are not reported.

Correlations between constructs were estimated by first calculating rater, professor, and relationship scores following the examples of Cook and Kenny (2004); Kwan, John, Kenny, Bond, and Robins (2004); and Kwan, John, Robins, and Kuang (2008). Professor scores were simply the mean score of each professor averaged across raters and items ($N = 10$). Rater scores were simply the mean score of each rater averaged across professors and items ($N = 145$). Relationship scores ($N = 483$) were calculated using the formula $\text{Rel}_{ij} = X_{ij} - \text{MRater}_i - \text{MProfessor}_j$, for which X_{ij} indicates a score on a variable for rater i and professor j, MRater_i is rater i's mean score across professors, and MProfessor_j is professor j's mean score across raters. Relationship scores were averaged across items. We did not adjust for section in calculating these scores as there were no effects for section on the study variables. Once rater, professor, and relationship scores were calculated, we used conventional correlation and regression analysis. We used percentile bootstrapping with 1,000 resamples to estimate statistical significance for correlations based on professor and relationship scores as these scores violated the independence of observations assumption. We used parametric significant tests for rater scores.

In our team's previous analyses (e.g., Gross *et al.*, 2009), we estimated multivariate generalizability correlations (Cronbach *et al.*, 1972) between constructs using the software *Mgenova* (Brennan, 2001). Although appropriate for the study's design, *Mgenova* has two disadvantages. Statistical control is cumbersome, and significance tests must be bootstrapped by hand. Fortunately, the results of the analyses just described and the results from *Mgenova* yielded identical results.[2]

Results

First, we attempted to replicate Gross *et al.*'s (2009) findings of large relationship, professor, and rater effects on teaching evaluations and on affect, as well as relationship effects on quiz performance.

As predicted, there were large relationship effects for teaching evaluations for both trailers and live lectures, with each accounting for about 40% of the variance (Table 1). That is, some professors elicited unusually favourable evaluations from some raters, more favourable than how the rater typically evaluated professors and more favourable than how the professor was typically evaluated by others. There were

[2] *We estimated the correspondence between the results of the two approaches by intraclass correlation for absolute agreement. The correlations produced by the two approaches were the dependent variables. Approach was one factor (Mgenova vs. the current approach), and variable pair was the second factor (e.g., positive affect and teaching evaluation). This design indicates the extent to which the two approaches yielded the same results. The correspondence between the results obtained by the two approaches was .98.*

Table 1. Variance components, standard errors, and effect sizes for study variables

	Variance component	Standard error	Proportion of variance explained
Teaching evaluations (trailer)			
Rater	.105	.022	.182*
Professor	.201	.099	.347*
Relationship	.226	.019	.392*
Positive affect (trailer)			
Rater	.318	.051	.378*
Professor	.156	.079	.185*
Relationship	.264	.023	.314*
Negative affect (trailer)			
Rater	.091	.018	.306*
Professor	.000	.002	.001
Relationship	.157	.014	.527*
Teaching evaluations (class)			
Rater	.098	.021	.184*
Professor	.191	.093	.359*
Relationship	.215	.018	.404*
Positive affect (class)			
Rater	.366	.060	.351*
Professor	.181	.092	.174*
Relationship	.369	.032	.354*
Negative affect (class)			
Rater	.090	.015	.418*
Professor	.003	.003	.014
Relationship	.070	.007	.324*
Quiz (class)			
Rater	.009	.002	.148*
Professor	.008	.006	.123
Relationship	.008	.002	.115*

Note. *$p < .05$. The variance components for all variables for the section factor were zero and not significant.

also large relationship effects for both positive and negative affect, with each accounting for more than 30% of the variance. That is, some professors elicited unusually favourable affect in some raters, but not others. There were also significant relationship effects on quiz scores. Some professors elicited unusually good memory for lectures in some students, but not others.

There was a substantial professor effect (i.e., inter-rater agreement) on teaching effectiveness for both trailers and live lectures, accounting for about 35% of the variance (Table 1). That is, raters agreed to a large extent that some professors were more effective than others. In addition, some professors consistently elicited more positive affect in raters than did other professors. There were no professor effects for negative affect or quiz scores.

There were large rater effects for teaching evaluations for both trailers and live lectures, accounting for nearly 20% of the variance (Table 1). That is, some raters consistently gave professors high scores, and other raters consistently gave professors low scores. Similarly, there were large rater effects for positive and negative affect for both

trailers and live lectures, accounting for more than 35% of the variance. Some raters consistently reported high positive or low negative affect across professors and time. Finally, there were significant rater effects for quizzes. Some students had consistently higher scores across quizzes than did other students.

Our primary question was, could we forecast from teaching trailers the students who responded unusually well to specific professors' live lectures (i.e., relationship effects)? As predicted, relational teaching evaluations and positive affect in response to trailers significantly forecasted evaluations of live lectures (Table 2). That is, when a student rated a professor's trailer unusually favourably, or experienced unusually high positive affect, the student also rated the professor's live lecture unusually well. In multiple regression analyses, teaching evaluations and positive affect in response to the trailers forecasted 7% (R = .27) of the variance in live lectures. Positive affect uniquely forecasted ratings of live lectures (β = .18; $p < .05$; ΔR^2 = .02), but evaluations of trailers did not (β = .11; n.s.; ΔR^2 = .01). The small predictive accuracy of the unique predictors (compared with the full equation) shows that most of the predictive accuracy was shared between evaluations and positive affect. Negative affect during trailers did not forecast evaluations of live lectures.

Table 2. Correlations between constructs for relationship, professor, and rater effects

	Positive affect (trailer)	Negative affect (trailer)	Teaching evaluations (class)	Positive affect (class)	Negative affect (class)	Quiz (class)
Teaching evaluations (trailers)						
Rater	.55*	−.03	.60*	.39*	−.08	.17*
Professor	.98*	NC	.86*	.74*	NC	NC
Relationship	.72*	−.22*	.24*	.23*	−.09*	.08
Positive affect (trailers)						
Rater	–	.34*	.36*	.76*	.16*	.14
Professor	–	NC	.87*	.76*	NC	NC
Relationship	–	−.15*	.26*	.30*	−.07	.08
Negative affect (trailers)						
Rater		–	.00	.34*	.34*	−.09
Professor		–	NC	NC	NC	NC
Relationship		–	−.03	−.03	−.03	−.08
Teaching evaluations (class)						
Rater			–	.51*	−.13	.19*
Professor			–	.94*	NC	NC
Relationship			–	.69*	−.22*	.16*
Positive affect (class)						
Rater				–	.33*	.08
Professor				–	NC	NC
Relationship				–	−.18*	.12*
Negative affect (class)						
Rater					–	−.09
Professor					–	NC
Relationship					–	−.02

Note. *$p < .05$. N = 145 for rater correlations, N = 10 for professor correlations, and N = 483 for relationship correlations. NC, not calculated as at least one of the variance components was not significant.

For professor effects, teaching evaluations and positive affect in response to trailers predicted evaluations of live lectures with superb accuracy (Table 2). That is, the professors who elicited consensus favourable evaluations and positive affect in trailers also elicited consensus favourable evaluations of live lectures. There were no significant professor effects for negative affect for either trailers or live lectures, and so, negative affect was not used in prediction. Evaluations and positive affect in response to trailers were nearly perfectly correlated. That is, for trailers, the consensus about teaching quality was the same empirically as the consensus about elicited positive affect. Multiple regression indicated that evaluations and positive affect from trailers predicted 76% of the variance in evaluations of live lectures (R = .87). Given how highly intercorrelated the two predictors were, and the limited sample size of professors, we did not estimate which predictor had unique predictive power.

Although not the primary focus of our research, we report predictive accuracy for rater effects for the sake of completeness. For rater effects, we could forecast evaluations of live lectures with excellent accuracy from evaluations and positive affect in response to trailers (Table 2). That is, raters who typically evaluated trailers favourably also evaluated live lectures favourably. Raters who typically responded to trailers with positive affect also typically evaluated live lectures favourably. Multiple regression analyses showed that evaluations of trailers predicted evaluations of live lectures significantly (β = .58; ΔR^2 = .24; t = 7.24; $p < .05$), but positive affect did not (β = .04; ΔR^2 = .00; t = 0.47; n.s.). The two constructs together forecasted 36% of the variance in teaching evaluations (R = .60). In contrast, negative affect in response to trailers did not forecast evaluations of live lectures. Raters who typically evaluated teaching trailers positively also had higher scores on the quizzes.

Finally, there were a number of interesting cross-sectional findings (Table 2). Perhaps most important, relationship effects on teaching evaluations and positive affect for live lectures were linked to relational quiz scores. That is, when a student saw a professor's lecture as unusually effective, or the professor elicited unusually high positive affect in a student, the student also scored especially well on the corresponding quiz. Other effects included that teaching evaluations were strongly linked to positive affect for each of the three effects for trailers and live lectures. For professor effects, the link between teaching evaluations and positive affect indicated that inter-rater agreement about which professors were more effective overlapped substantially with inter-rater agreement about which professors elicited more positive affect. For relationship effects, when a rater evaluated a professor unusually favourably, the professor also elicited unusual favourable affect. For rater effects, raters who typically responded to all professors with positive affect also typically evaluated all professors favourably. In contrast, negative affect was linked to poor evaluations only for relationship effects. When a professor elicited unusually high negative affect, the rater saw the professor as unusually ineffective, for both trailers and live lectures.

Discussion

We could forecast the professor that a student would see as unusually effective in a live lecture from the student's reactions to 6-min video trailers of the professor (i.e., relationship effects). We could also forecast from trailers which professors would be seen as more effective (on average across students) than other professors in live lectures (i.e., professor effects). These findings support the possibility of developing online systems

that would provide personalized recommendations that specific students take courses from specific professors.[3] This matching has the potential to improve teaching effectiveness and student satisfaction. It is important to forecast both relationship and professor effects, because each has a distinct link to student learning. In the present study and Gross *et al.*'s (2009) studies 2 and 3, relational memory for lecture was linked to relational student evaluations of teaching effectiveness. That is, when a student evaluated a professor's lecture as unusually effective, the student remembered an unusually large amount about the lecture. It is already well established that professors with higher teaching evaluations averaged across students (i.e., professor effects) elicit more student learning (Cohen, 1981; Feldman, 1989a; Marsh, 1984, 2007).

That we could forecast relationship effects for teaching effectiveness is consistent with the research on forecasting relational perceived support. Relationship effects are especially large in perceived support, and thus, support interventions might be improved if support providers could be matched to support recipients such that unusually supportive relationships emerged. Veenstra *et al.* (2011) investigated the predictive accuracy of (1) very brief conversations between recipients and providers and (2) recipients' reactions to short video interviews of providers. In social support interventions, it would be more cost effective for recipients to view recorded interviews of providers rather than to have brief conversations with every potential provider. Veenstra *et al.* (2011) could forecast relational perceived support several months in advance from brief conversations, but not from video interviews.

The accuracy of forecasting relational teaching evaluations was comparable to forecasting job performance from conscientiousness – one of the strongest predictors of job performance (Oh, Wang, & Mount, 2011). Nonetheless, the accuracy of relational forecasting was small compared to forecasting based on professor and rater effects. Veenstra *et al.* (2011) also found that rater effects were more predictable than relationship effects. We suspect these differences in predictability result because professor and rater effects have stability as part of their operational definitions. By definition, professor effects are stable across raters and rater effects are stable across professors. In contrast, relationship effects are unstable across professors and raters.

The very high predictive accuracy of professor effects was predicted by Kenny's (2004) quantitative theory of person perception (PERSON) and is consistent with Park *et al.*'s (1997) study of the longitudinal stability of target effects. According to PERSON, inter-rater agreement reaches asymptote very quickly as more information about targets is presented. Agreement is driven by shared stereotypes initially (e.g., extroverted professors are better), but agreement based on targets' actions quickly dominates. Thus, it takes surprisingly few observations to achieve asymptotic agreement, and the video trailers apparently achieved this.

Our findings are also consistent with 'thin slices' research (Ambady & Rosenthal, 1992, 1993; Babad, Avni-Babad, & Rosenthal, 2004), in which raters' evaluations of brief (<5 min) exposure to targets forecasted information derived from other sources (e.g., teaching evaluations). Thin slices research on teaching reflects professor effects, because

[3] *In recommending from which professor a student should take a class, one would use a linear equation that forecasted the student's evaluation of each professor from the student's average score, each professor's average score, and the relationship score for each student–professor dyad. Typically, recommendations would be made separately for each student. In this case, the student's average score can be dropped from the equation because the score would be the same for predicting reactions to each professor. When a professor's standardized score is zero, prediction would be based entirely on the relationship score. When the relationship standardized score is zero, prediction would be based entirely on the professor score.*

professors' scores are averaged across raters and professors are the unit of analysis (Kenny, 2004). Our ability to forecast professor effects was much more precise (r = .86) than obtained in the thin slices research, where the meta-analytic estimate was r = .41 (Ambady & Rosenthal, 1992). According to Kenny's (2004) PERSON model, the stronger predictive accuracy for professors in the current study likely reflected that the same raters viewed both trailers and lectures, whereas in most thin slices research, teaching evaluations and reactions to thin slices are provided by different raters. According to PERSON, the stability of professor effects is suppressed when raters have access to different information. The impressive predictive accuracy for professor scores in the current study might have also resulted from the great care taken in constructing the trailers to maximize the representativeness of professors' teaching.

Our findings of strong and significant rater, professor, and relationship effects for teaching evaluations replicate previous studies (Gross *et al.*, 2009) and have implications for interpreting teaching evaluations in educational settings. Teaching evaluations are typically interpreted as though they reflect only professor effects (i.e., the extent to which some professors are better than others). Yet, only a portion of teaching evaluations reflects this effect. Other important influences are rater bias and relationship effects. The presence of strong rater and relationship effects opens the door for decision errors when teaching evaluations are treated as though they reflect only professor effects. For example, if the tendency of some students to rate all professors favourably (rater bias) is correlated with major, then professors in certain departments will have inflated teaching evaluations. Relationship effects could also bias teaching evaluations. If a professor is better at teaching advanced students than introductory students, he or she might have unusually low scores if assigned to teach many sections of introductory courses. The variance partitioning approach integral to generalizability theory (Cronbach *et al.*, 1972) and the SRM (Kenny, 1994; Kenny *et al.*, 2006) informs personnel decisions by showing that professors' scores on teaching evaluations do not primarily reflect the simple case that some professors are better than others.

The current research replicated previous findings (Fortunato & Mincy, 2003; Gross *et al.*, 2009) that teaching evaluations are strongly linked to positive affect. That is, students rate professors as effective insofar as professors elicit positive affect. Yet, the psychological mechanisms differ for relationship, professor, and rater effects. Relationship effects reflect the unique match between specific professors and students. When a professor elicits unusually positive affect in a student, that student rates the professor as unusually effective. Rater effects reflect trait like personality processes in that the students who characteristically experience positive affect also characteristically rate professors as effective. Professor effects reflect the trait like personality characteristics of professors. Professors who consistently elicit positive affect in students also consistently elicit favourable teaching evaluations.

We interpret findings for relationship effects on teaching evaluations, affect, and memory in terms of relational regulation theory (Lakey & Orehek, 2011). According to this theory, social interaction is a key way by which people regulate their own affect and cognition, and this regulation is largely relational. From this perspective, some professors are unusually effective in regulating some students' positive affect and memory and are rewarded with high teaching evaluations. This does not mean, however, that eliciting positive affect is not an important part of teaching, as positive affect includes attentiveness and interest.

We should note some limitations of the current study. First, we studied only psychology lectures and the results might not generalize to other fields. Second, we studied only 40-min lectures and the results might not generalize to semester-long courses. It will be important to

determine whether responses to trailers will forecast teaching evaluations based on entire-semester courses. This is especially important as teaching evaluations might be based, in part, on interactions with the professor outside of the classroom (e.g., office hours); the teaching trailers developed for the present study did not reflect extra-classroom behaviour. Third, preparing carefully constructed trailers as in the present study might be too expensive to justify the cost-effectiveness of forecasting students' responses to teaching. It will be important to determine the extent to which very small slices (Ambady & Rosenthal, 1992) of professors' actions can forecast students' relational responses to specific professors. Fourth, professors were confounded with topic. For example, one professor spoke about psychopathology and another lectured on text comprehension. Thus, relational effects might reflect students' reactions to different topics as much as reactions to different professors. However, it would be difficult to disentangle student × professor interactions (i.e., relationship effects) from student × topic interactions. To do so, one would need a design in which professors and topics were fully crossed and each professor taught each topic. However, having students hear different faculty members give the same lecture would be sufficiently strange as to present another set of interpretive problems. Fifth, our statistical procedures require no missing data, and thus, the small numbers of students who missed a session were excluded from analyses. It is possible that students who missed sessions differed in some important way from students who attended all sessions. For example, students lower in conscientiousness might be under-represented in the data and such students might not show the same findings as the full sample. Sixth, memory for lecture was assessed with multiple-choice questions in the next class session. Responses to such questions might not generalize to performance on essay examinations or papers and might not last beyond a few days. Finally, the current results need replication.

In conclusion, it appears possible to forecast from brief video trailers, the extent to which an individual student will respond unusually well to a specific professor, as well as how students on average will respond to different professors. These findings suggest the feasibility of developing online systems that would recommend that a student take courses from specific professors. Such an approach might improve student satisfaction and the effectiveness of instruction.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*, 256–274. doi:10.1037/0033-2909.111.2.256

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, *64*, 431–441.

Babad, E., Avni-Babad, D., & Rosenthal, R. (2004). Prediction of students' evaluations from brief instances of professors' nonverbal behavior in defined instructional situations. *Social Psychology of Education*, 7, 3–33. doi:10.1023/B:SPOE.0000010672.97522.c5

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Centra, J. A. (1975). Colleagues as raters of classroom instruction. *Journal of Higher Education*, *46*, 327–337. doi:10.2307/1980806

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, *51*, 281–309. doi:10.3102/00346543051003281

Cook, W. L. (2000). Understanding attachment security in family context. *Journal of Personality and Social Psychology*, *78*, 285–294. doi:10.1037/0022-3514.78.2.285

Cook, W. L., Kenny, D. A., & Goldstein, M. J. (1991). Parental affective style risk and the family system: A social relations model analysis. *Journal of Abnormal Psychology*, *100*, 492–501. doi:10.1037/0021-843X.100.4.492

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley & Sons.

Feldman, K. A. (1989a). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*, 583–645. doi:10.1007/BF00992392

Feldman, K. A. (1989b). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, *30*, 137–194. doi:10.1007/BF00992716

Fortunato, V. J., & Mincy, M. D. (2003). The interactive effects of dispositional affectivity, sex, and a positive mood induction on student evaluations of teachers. *Journal of Applied Social Psychology*, *33*, 1945–1972. doi:10.1111/j.1559-1816.2003.tb02088.x

Gross, J. A., Lakey, B., Edinger, K., Orehek, E., & Heffron, D. (2009). Person perception in the college classroom: Accounting for tastes in students' evaluations of teaching effectiveness. *Journal of Applied Social Psychology*, *39*, 1609–1638. doi:10.1111/j.1559-1816.2009.00497.x

Kenny, D. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford.

Kenny, D. A. (2004). PERSON: A general model of interpersonal perception. *Personality and Social Psychology Review*, *8*, 265–280. doi:10.1207/s15327957pspr0803_3

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Analysis of dyadic data*. New York, NY: Guilford Press.

Kwan, V. S. Y., John, O. P., Kenny, D. A., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review*, *111*, 94–110. doi:10.1037/0033-295X.111.1.94

Kwan, V. S. Y., John, O. P., Robins, R. W., & Kuang, L. L. (2008). Conceptualizing and assessing self-enhancement bias: A componential approach. *Journal of Personality and Social Psychology*, *94*, 1062–1077. doi:10.1037/0022-3514.94.6.1062

Lakey, B., Cohen, J. L., & Neely, L. C. (2008). Perceived support and relational influences on psychotherapy process constructs. *Journal of Counseling Psychology*, *55*, 209–220. doi:10.1037/0022-0167.55.2.209

Lakey, B., & Orehek, E. (2011). Relational Regulation Theory: A new approach to explain the link between perceived support and mental health. *Psychological Review*, *118*, 482–495. doi:10.1037/a0023477

Livi, S., Kenny, D. A., Albright, L., & Pierro, A. (2008). A social relations analysis of leadership. *The Leadership Quarterly*, *19*, 235–248. doi:10.1016/j.leaqua.2008.01.003

Lucas, T., Lakey, B., Arnetz, J. E., & Arnetz, B. B. (2010). Do cultural competency ratings reflect characteristics of providers or perceivers? Initial demonstration of a generalizability theory approach. *Psychology, Health & Medicine*, *15*, 445–453. doi:10.1080/13548506.2010.482141

Marsh, H. M. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77–95.

Marsh, H. M. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707–754. doi:10.1037/0022-0663.76.5.707

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

Oh, I., Wang, G., & Mount, M. K. (2011). Validity of observer ratings of the five-factor model of personality traits: A meta-analysis. *Journal of Applied Psychology*, *96*, 762–773. doi:10.1037/a0021832

Park, B., Kraus, S., & Ryan, C. S. (1997). Longitudinal changes in consensus as a function of acquaintance and agreement in liking. *Journal of Personality and Social Psychology*, *72*, 604–616. doi:10.1037/0022-3514.72.3.604

Veenstra, A., Lakey, B., Cohen, J. L., Neely, L. C., Orehek, E., Barry, R., & Abeare, C. (2011). Forecasting the specific providers that recipients will perceive as unusually supportive. *Personal Relationships*, *18*, 677–696. doi:10.1111/j.1475-6811.2010.01340.x

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. doi:10.1037/0022-3514.54.6.1063

Received 30 July 2013; revised version received 19 May 2014