Research Journal of Education
ISSN(e): 2413-0540, ISSN(p): 2413-8886
Vol. 8, Issue. 3, pp: 70-79, 2022
URL: https://arpgweb.com/journal/journal/15
DOI: https://doi.org/10.32861/rje.83.70.79

Academic Research Publishing Group

Original Research **Open Access**

Finding a Simple Solution to the Problem of Student Evaluations: An Index of Traditional Evaluation Questions

Derek Stimel (Corresponding Author)
Ph.D. Associate Professor of Teaching Economics Department of Economics University of California at Davis One Shields Avenue Davis, CA 95616
Email: dstimel@ucdavis.edu

Janine Lynn Flathmann Wilson
Ph.D. Associate Professor of Teaching Economics Department of Economics University of California at Davis One Shields Avenue Davis, CA 95616

Article History
Received: 27 May, 2022
Revised: 23 July, 2022
Accepted: 17 August, 2022
Published: 23 August, 2022

Abstract

Literature commonly finds student evaluations of economics instruction do not assess true teaching effectiveness. Many techniques improving learning outcomes result in lower student evaluations because students must work harder. Still, we often rely on a single overall rating for assessing teacher quality. We construct a model of student responses that indicate likeability of the instructor and difficulty of the instruction. We test the predictive power of these indicators using teaching evaluations of intermediate microeconomics instructors over five years alongside data on student grades in future courses that depend on intermediate microeconomics. Using these questions to construct a more meaningful evaluation of how well the intermediate microeconomic instructors prepare students to recall and use important concepts in the future improves our ability to evaluate instructor effectiveness.

Keywords: Student evaluations; Instructor quality.

1. Introduction

The goal to provide more effective teaching to our undergraduate students becomes more important with the rising costs of a college education. Students are no longer content to sit and observe a brilliant professor perform a monologue. Instead, students expect to receive effective instruction. Instructors that can give a student the ability to understand material deeply and retrieve this information months or years into the future provides value for the money spent on higher education. As we adapt to this environment, as administrators and instructors, with the goal to provide effective instruction, it becomes necessary to create evaluation methods that capture how well an instructor is teaching.

Often, a faculty member under review for tenure will be assessed using student completed teaching evaluations. For many of us, a single question is used: "How effective was the instructor?" or "How well did the instructor teach the course?" There is a strong incentive for a tenure track faculty member to please the students during the quarter in order to inspire them to answer "YES, my instructor was amazing". However, pleasing the students is not the goal of effective teaching. Effective teaching requires that faculty engage the students in order that they can use the tools that we have given them far into the future. We explore a method to choose student evaluation questions in order to assess and then reward high quality instruction within the confines of a university with limited resources.

The remaining paper is organized in the following way. Section 2 discusses the literature findings on how teaching quality impacts learning and the evaluation of this quality. Section 3 summarizes and describes our student performance and instructor evaluation data. Section 4 is where we group teaching evaluation responses into those that evaluate likability and those that evaluate course difficulty. Section 5 is where we do the main regression analysis and establish the predictability of perceived instructor likability and perceived course difficulty. Finally, in section 6 we conclude with a description of the application of our findings and future avenues for research.

2. Literature Review

This idea that an instructor can add value to student learning and retention of course material has been argued by a wide variety of papers. Aaronson *et al.* (2007) use public high school data to establish a clear link between instructor quality and student outcomes, finding a significant impact of teacher quality on math proficiency. Work done at the post-secondary level finds a positive impact of instruction on student scores (Jonah and Rockoff, 2004; Kane *et al.*, 2008) in addition to lowering dropout rates among students (Florian *et al.*, 2009). Rivkin *et al.* (2005) go further to compare higher levels of teacher quality with the advantages of lower class sizes and suggest that a focus on teacher quality improvements would produce more positive student outcomes than lowering class sizes. Rivkin *et al.* (2005) and Thomas (2011) also suggest that we cannot measure teaching quality using observable

characteristics. Years of experience and teacher salary are not useful when predicting teacher effectiveness in the classroom. Though Brian *et al.* (2004) find that principal evaluations of teachers are the best way to evaluate instructor quality and Charles *et al.* (2006) advocate the use of National Board Certification and teacher licensure test scores to predict teacher quality, at the post-secondary level of instruction Kane *et al.* (2011) find that using external, highly trained evaluators is most effective in measuring teacher quality. Though this is interesting from a theoretical perspective, many schools and colleges do not have the resources to use objective instructor evaluations with highly trained in-class reviewers on a regular basis.

We are often left only with the option to evaluate instructors with subjective student evaluations. Bruce *et al.* (2010), warns us that these methods can be predictive of student success only for a student's current grade in a course and not in follow-on courses. Their concern is that a student does not truly know what they learned in a class. Without this understanding, they see an instructor for their current classroom experience without understanding how the course has prepared them for the future. Scott *et al.* (2010), also find that contemporaneous student achievement is predicted by student evaluations but has little use in follow-on courses. In fact, they found that instructors that get higher marks on student evaluations harm the student follow-on achievement in more advanced courses.

In addition to our concern that subjective student evaluations do not predict retention of knowledge are the biases present in subjective student evaluations. Lillian *et al.* (2014) and Anne and Boring (2017) have used the online course format to present the same course to students under a male instructor and a female instructor. They find that student evaluations are consistently higher for the male instructors. These gender biases have dangerous implications for schools using these subjective student evaluations to make tenure decisions.

The discipline continues to use subjective student evaluations because they are simple and cost-effective. Expert observations in classrooms are costly and cumbersome. Our research attempts to use the an index of subjective student evaluation responses to gather more relevant information about instructor quality.

Our paper argues that we need to be more thoughtful about the student evaluation questions that we use to assess instructor effectiveness. We break down subjective teaching evaluation questions into two groups: instructor likability and instructor difficulty. Testing these two aspects of student impressions of instructors gives us a greater ability to predict the long-term effectiveness of instruction. We find that when a student considers a course more difficult, they will be more likely to effectively use that knowledge in future courses.

3. Data and Descriptive Statistics

Our data is from a large public university in the western US. We have data for students connected to intermediate microeconomics sections taught between fall 2011 and spring 2016 quarters including some summer quarters. There are total of 40 sections of the course over that time taught by 13 different instructors. The data comes from two separate sources.

First we have the anonymous teaching evaluation data completed by students during the last week of the quarter. This data cannot be linked to specific students as it is fundamentally anonymous for each course. Students are provided statements about the course and its instruction and then asked the extent they agree or not with those statements on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The evaluation had 10 statements up until winter 2014 quarter when two additional statements were added, making the total 12. Students are not required to complete the evaluation and are also allowed to skip statements if they desire. Additionally the evaluations contain open-ended free response questions but we do not have that qualitative data.

Second, we have specific data about students that enrolled and completed intermediate microeconomics during this time period. We do not have data on attrition in the form of drops or withdraws, though anecdotally the attrition rate is typically low given intermediate microeconomics is a core required course for the students taking it and is not a general education course. For specific students we have their demographic information (e.g. gender, international student status, ethnicity, transfer student status, first generation status, low income status, language spoken at home), academic aptitude (e.g. SAT writing and math scores, the number of times they have taken intermediate microeconomics), grade earned in intermediate microeconomics, and their grade earned in the subsequent quarter for courses that have intermediate microeconomics as a prerequisite. Of that group we will ultimately drop students out of the certain specification for largely one of two reasons. One, some students do not have SAT scores listed, typically transfer students. We will retain those students in some specifications but they will be dropped in others. Second, some students took intermediate microeconomics or the follow-on courses pass/fail rather than letter grade (the latter is the more likely possibility of the two). As we are ultimately interested in the follow-on course performance, those students taking courses pass/fail will end up being dropped from the study entirely.

Table-1. Descriptive Statistics of Student Evaluations

Statement	Description	Number	Mean	Median	Standard Deviation
1	Instructor Clarity & Organization	2801	4.11	4.00	1.00
2	Instructional Value of Assignments	2793	4.23	5.00	0.96
3	Instructor Availability & Helpfulness	2351	4.34	5.00	0.83
4	Instructor Responsiveness to Difficulty	2756	4.23	4.00	0.90
5	Fair Exam Content	2800	4.05	4.00	1.07
6	Fair & Timely Grading	2788	4.26	5.00	0.92
7	Workload Appropriate to Units	2797	4.32	5.00	0.86
8	Intellectual Challenge of Course	2796	4.39	5.00	0.78
9	Stimulated Interest in Subject	2795	3.99	4.00	1.07
10	Instructor's Overall Teaching	2799	4.18	4.00	0.95
11	Instructor Overall Teaching Effectiveness	1453	4.07	4.00	0.99
12	Educational Value of Course	1452	4.17	4.00	0.89

Table 1 provides summary statistics of the teaching evaluation scores across all the intermediate microeconomics sections in our sample. A higher score indicates a more positive assessment by the student of that particular item. The mean response for each statement is around 4, suggesting that a favorable "agree" is the typical choice. Comparing the mean and median we can see some evidence of skewness in certain statements where the median response is "strongly agree" (i.e. 5) but the mean response is "agree" (i.e. 4). This likely indicates a relatively small percentage of students in those items giving very low scores (i.e. a 1 or 2).

We would expect responses by students on these statements to be positively correlated across the statements. Students that have a positive view of the course or instructor would tend to give high marks across the board and those with negative views of the course or instructor would tend to give low marks across the board. Table 2 provides the correlations among the statements. Statements 10, 11, and 12 are highly correlated. That is not surprising as it is hard to tell what the precise difference is between those statement (particularly 10 and 11) and it's not clear if students would see much difference between them. Of note is that statement 1 and 9 are also relatively highly correlated with 10, 11, and 12 as well. That perhaps suggests that the instructor's organization and clarity in the course along with the ability to stimulate interest in the course subject matter may be of particular importance to a student's overall rating.

Table-2. Correlations of Student Evaluations

Statement	1	2	3	4	5	6	7	8	9	10	11	12
1	--											
2	0.51	--										
3	0.49	0.53	--									
4	0.57	0.55	0.64	--								
5	0.25	0.49	0.35	0.44	--							
6	0.44	0.53	0.52	0.54	0.53	--						
7	0.40	0.56	0.50	0.53	0.57	0.60	--					
8	0.43	0.43	0.47	0.47	0.24	0.44	0.49	--				
9	0.58	0.53	0.49	0.58	0.45	0.50	0.52	0.51	--			
10	0.79	0.62	0.59	0.67	0.43	0.56	0.55	0.49	0.67	--		
11	0.78	0.60	0.55	0.65	0.45	0.53	0.55	0.47	0.66	0.88	--	
12	0.67	0.60	0.53	0.60	0.45	0.53	0.60	0.56	0.76	0.75	0.78	--

For the student specific data it is important to mention that we will have some duplication of students in the sample. That is, if a student took two follow-on courses in the subsequent quarter that had intermediate microeconomics as a prerequisite they would enter in the data as two separate observations. Thus for the student specific entries we have a maximum of 4304 entries but only 3253 students. Most students only took one follow-on course in the subsequent quarter but some took 2 and a very limited number took 3 or 4. Due to the aforementioned issues of SAT scores and pass/fail grades and a few other minor issues, the effective samples for the regressions are substantially smaller than either of those numbers.

The economics department in the sample has a grading policy that essentially strives to set the mean grade in all courses at a B- (2.7 grade point average). For the intermediate microeconomics courses we have overall mean grade

point average (GPA) of 2.60, median of 2.70, and standard deviation of 0.96. For the follow-on courses we have an overall mean GPA of 2.67, median of 2.70, and standard deviation of 0.99. We can see in aggregate then that the grades being issued conform to the departmental standards.

4. Preliminary Analysis

Our prior belief based on our own experiences as instructors was that student evaluations are a function of two underlying unobserved variables: instructor likability and instructor difficulty. As instructors ourselves, our preconceived perception is that high marks on evaluation statements such as 1,3,4,9,10,11,12 (see Table 1) are driven by likability and statements 2,5,6,7,8 are driven by difficulty. To assess this, we started by conducting an exploratory factor analysis following the recommendations for applying this methodology to social sciences found in Costello & Osborne (2005). We did the same analysis for the first ten statements which are available on all evaluations and then a sub-sample of the twelve statements that were available only in winter 2014 onward. We used maximum likelihood estimation of the factors, scree plots to decide the number of factors, and oblique rotation using the oblimin method. We also report the eigenvalues for the selected factors. To decide the loadings, we simply assigned each statement to the factor with the largest rotated loading. A traditional cutoff such as 0.30 for loadings would create some ambiguity for statements 4,8, and 9; for the other statements, the loadings are relatively clear.

Figures 1A and 1B show the Scree plots for the two factor analysis. Each scree plot suggests two factors in the evaluation data.

Figure-1A

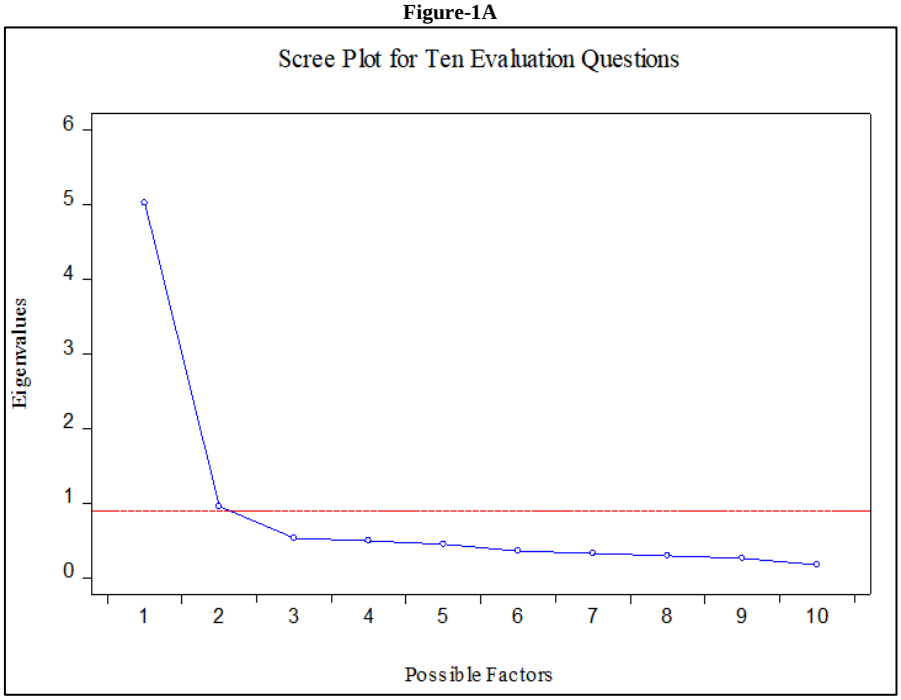

Figure 1B

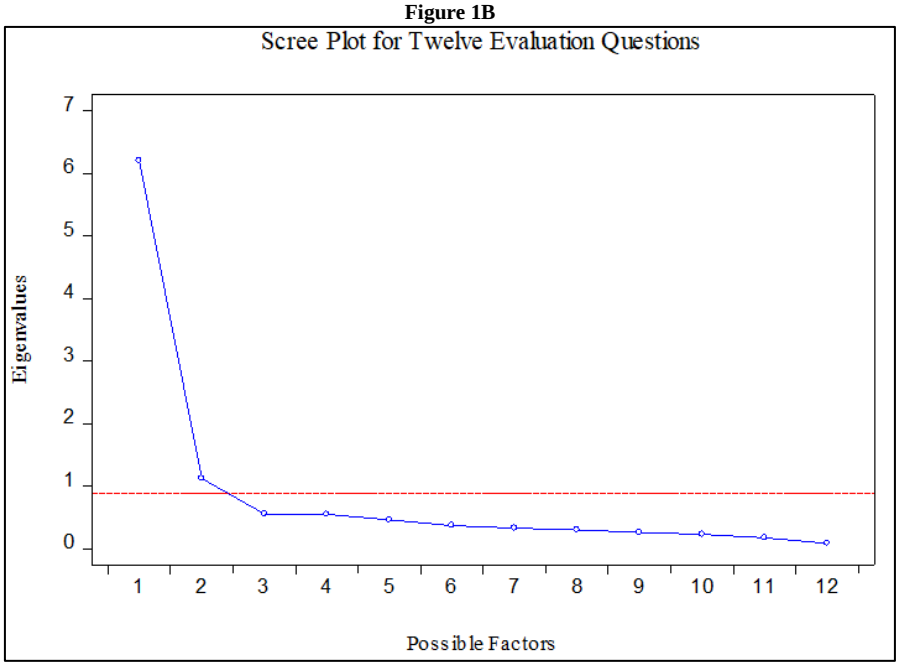

Table 3 shows the factor loadings for the two exploratory factor analyses and the reported eigenvalues for the selected factors. While the scree plots do suggest two factors, if we went by the commonly used criteria of factors with eigenvalues greater than one, the second factor when ten statements are used would not be selected though its eigenvalue is close to one. As we can also see in Table 3 there is some disagreement about where statement 8 would load and statements 4 and 9 load on both factors using solely the 0.3 or greater standard. Statement 8 about intellectual challenge of the course is particularly interesting in that we thought it was the most likely statement associated with course difficulty and the factor analysis suggests that may not be the case. Returning to our prior beliefs, the factor analysis does confirm some of our hypothesized split but not fully. Of course it is not possible to definitively say whether these two identified factors in fact represent likability and difficulty. If we do fit these factors to that dichotomous split, it appears that factor 1 is the likability factor as statements 1,9,10,11,12 were all hypothesized to be part of likability and factor two would be difficulty. We next constructed factor variables based on the splits identified in Table 3. An alternative possibility is that some element of likability and difficult of the instructor drives each and every response to the evaluation statements and splitting the statements between those two latent factors is too restrictive. Thus we also constructed factor variables using regression to obtain the estimated optimal factor scores and creating two weighted factor indices that rely on all statements.

Table-3. Factor Analysis

		Ten Statements		Twelve Statements	
Statement	**Description**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
1	Instructor Clarity & Organization	**0.94**	-0.10	**0.89**	-0.07
2	Instructional Value of Assignments	0.21	**0.53**	0.19	**0.54**
3	Instructor Availability & Helpfulness	0.23	**0.41**	0.11	**0.50**
4	Instructor Responsiveness to Difficulty	0.37	**0.43**	0.30	**0.47**
5	Fair Exam Content	-0.11	**0.88**	-0.10	**0.84**
6	Fair & Timely Grading	0.10	**0.64**	0.04	**0.68**
7	Workload Appropriate to Units	0.07	**0.61**	0.01	**0.65**
8	Intellectual Challenge of Course	**0.27**	0.22	0.17	**0.31**
9	Stimulated Interest in Subject	**0.50**	0.40	**0.48**	0.42
10	Instructor's Overall Teaching	**0.72**	0.22	**0.77**	0.17
11	Instructor Overall Teaching Effectiveness	---	---	**0.99**	-0.05
12	Educational Value of Course	---	---	**0.59**	0.18
	Eigenvalues	5.02	0.95	6.21	1.12

5. Main Analysis

As we now turn to examining how the student evaluations affect follow-on course performance, our dependent variable of interest is the grade earned in elective courses taken in the quarter subsequent to intermediate microeconomics. Given the ordinal discrete nature of that variable we use an ordered multinomial logit model. To do that, we recoded the grade variables for both the follow-on course and for the intermediate micro course from their usual 0-4 point scale to a 0-11 point scale (where 0 is F, 10 would be A-, 11 would be A/A+, etc.). Also at this level of analysis, we only can use the mean course responses on the evaluations as we do not have the ability to link the individual evaluations to specific students. Further we lose observations due to transfer student lacking SAT scores and inclusion of statements 11 and 12 that only began in winter 2014. Thus we examine some specifications that leave out those variables. In the case of the SAT scores, when they are left out of a specification we are then counting on the grade earned in intermediate microeconomics to be the proxy for academic aptitude. We include an indicator variable for summer school courses. We also explored an indicator variable for part-time faculty but due to the fact that all the summer courses are taught by part-time faculty, we dropped that variable due to collinearity problems. We are essentially treating an instructor teaching one course in our sample as different from the same instructor teaching the course again in our sample. To control for that issue we also include individual instructor fixed effects in some specifications. We encounter a collinearity problem between the summer school variable and the instructor fixed effects when statement 11 and 12 are included as there are simply not enough observations for sufficient distinction between the instructor fixed effects and the summer school indicator variable (i.e. it captures a clear split of instructors that only teach in summer vs. those that don't). Thus we left out the summer school indicator in those specifications.

Before we examined the explanatory power of the identified factors of likability and difficulty from our preliminary analysis we wanted to assess the explanatory power of the evaluation statements individually along with some of our other variables. Table 4A shows the initial full multinomial logit results where the model is estimated by maximum likelihood with Huber-White standard errors. Dashed lines indicate that a variable was left out of that particular specification. We will subsequently present more parsimonious models so for now we only present the independent variable coefficients and p-values leaving out the estimated dependent variable limits. The coefficients in Table 4 are unadjusted, thus we can only interpret the signs of the coefficients at this point, not the magnitudes of the coefficients.

Table-4A. Initial Multinomial Logit Results

Independent Variable	Coefficient Estimates [p-value]			
Instructor Clarity & Organization	0.739	0.430	9.756	4.878
	[0.52]	[0.53]	[0.27]	[0.36]
Instructional Value of Assignments	0.302	0.316	-3.702	-0.214
	[0.66]	[0.47]	[0.13]	[0.88]
Instructor Availability & Helpfulness	**2.334**	0.877	**9.308**	**4.682**
	[0.01]	[0.13]	[0.03]	[0.06]
Instructor Responsiveness to Difficulty	-1.155	-0.785	-3.839	-2.647
	[0.30]	[0.29]	[0.18]	[0.13]
Fair Exam Content	0.470	**0.763**	**7.542**	**5.114**
	[0.37]	[0.03]	[0.08]	[0.05]
Fair & Timely Grading	-0.516	-0.386	-2.838	-1.483
	[0.28]	[0.22]	[0.18]	[0.23]
Workload Appropriate to Units	0.436	0.425	1.052	-0.602
	[0.53]	[0.36]	[0.58]	[0.60]
Intellectual Challenge of Course	1.117	**1.408**	2.678	2.357
	[0.23]	[0.01]	[0.37]	[0.19]
Stimulated Interest in Subject	-1.043	-0.693	-7.091	-3.159
	[0.13]	[0.15]	[0.19]	[0.32]
Instructor's Overall Teaching	-0.456	-0.912	-1.935	**-4.126**
	[0.63]	[0.11]	[0.60]	[0.09]
Instructor Overall Teaching Effectiveness	---	---	-7.147	-1.200
			[0.40]	[0.80]
Educational Value of Course	---	---	3.497	0.194
			[0.62]	[0.96]
Intermediate Microeconomics Grade	**0.395**	**0.387**	**0.414**	**0.389**
	[0.00]	[0.00]	[0.00]	[0.00]
Gender	0.075	0.020	0.190	0.035
	[0.46]	[0.76]	[0.17]	[0.69]
Underrepresented Minority	-0.002	**-0.167**	-0.303	**-0.347**
	[0.99]	[0.09]	[0.12]	[0.01]
Low Income	**-0.305**	**-0.175**	-0.321	**-0.240**
	[0.01]	[0.03]	[0.19]	[0.02]
Language	0.072	**-0.089**	0.085	-0.027
	[0.27]	[0.02]	[0.33]	[0.61]
First Generation	-0.090	**-0.176**	-0.173	**-0.199**
	[0.46]	[0.00]	[0.32]	[0.04]
SAT Math	0.001	---	**0.003**	---
	[0.11]		[0.00]	
SAT Writing	**0.002**	---	**0.000**	---
	[0.01]		[0.73]	
Repeat	**-0.598**	**-0.956**	-0.130	**-0.938**
	[0.01]	[0.00]	[0.64]	[0.00]
Summer Course	**-1.706**	**-1.076**	---	---
	[0.01]	[0.01]		
Instructor Fixed Effects	Yes	Yes	Yes	Yes
Pseudo R^2	0.097	0.084	0.107	0.094
Observations	1480	3157	843	1803

In Table 4A the bold coefficients indicate variables that are statistically significant at a 10% level or better. Unsurprisingly the grade earned in intermediate microeconomics seems to be a good predictor of the grade earned in the follow-on course. That may be due to mastery of that prerequisite subject matter but may also simply be a proxy for general academic aptitude. The low income status of a student along with whether they've repeated a course or not seem to also be good predictors. For the evaluation statements Table 4A suggest instructor availability and helpfulness and fair exam content, both part of the instructor difficulty factor are good predictors of the follow-on course grade.

There are a relatively substantial of statistically insignificant variables in the specifications in Table 4A and a general concern about possible collinearity issues between the different evaluation statements. To address that we reduced the dimensions of the regressions in Table 4A using a general to specific methodology. Excluding the instructor fixed effects which we kept for all specifications, we began by removing the variable with the largest p-value (equivalently the lowest z-statistic) and re-estimating the model. We repeated that process of removing variables until all the variables remaining in the regression model with statistically significant at a 10% level (p-

value less than 0.1). Table 4B shows which variables we kept based on that methodology and is simply to provide further evidence on which variables may be good predictors of the follow-on course performance by students having taken intermediate microeconomics.

Table-4B. General to Specific Multinomial Logit Results

Independent Variable	Variable Retained?			
Instructor Clarity & Organization				Yes
Instructional Value of Assignments			Yes	
Instructor Availability & Helpfulness	Yes		Yes	
Instructor Responsiveness to Difficulty				Yes
Fair Exam Content		Yes		
Fair & Timely Grading				
Workload Appropriate to Units		Yes		Yes
Intellectual Challenge of Course		Yes		Yes
Stimulated Interest in Subject	Yes		Yes	
Instructor's Overall Teaching				Yes
Instructor Overall Teaching Effectiveness	---	---		
Educational Value of Course	---	---		
Intermediate Microeconomics Grade	Yes	Yes	Yes	Yes
Gender				
Underrepresented Minority		Yes		
Low Income	Yes	Yes	Yes	Yes
Language		Yes		
First Generation		Yes		Yes
SAT Math	Yes	---	Yes	---
SAT Writing	Yes	---		---
Repeat	Yes	Yes		Yes
Summer Course	Yes	Yes	---	---
Instructor Fixed Effects	Yes	Yes	Yes	Yes
Pseudo R^2	0.097	0.083	0.089	0.090
Observations	1547	3157	1553	3219

As removing variables allowed in some situations more observations to be added to a specification, the observations for the final specifications listed in Table 4B are larger than in Table 4A. Based on these results, the grade earned in intermediate microeconomics, low income status, and repeat course status appear to still be good predictors in the specifications. More evaluation statements show some evidence of being good predictors and the statement about fair exam content does not.

It is worth noting that both Table 4A and Table 4B show that the evaluation statements about overall teaching and overall teaching effectiveness, likely the evaluation statements focused on in an instructor's teaching ability in merit and tenure cases, are not particularly good predictors of student performance in their follow-on elective courses.

Having a sense now of what evaluation statements may or may not be good predictors of future student performance, we now estimate ordered multinomial models using the factors identified in the preliminary analysis. Reiterating, we looked at two sets of those factors: one set from the loadings identified in Table 3 and one set from regression estimated factor scores. Table 5A shows the coefficient estimates from those models.

Table-5A. Multinomial Logit with Factors

Factor Construction	Factors Based on Loadings				Factors Based on Regression			
Independent Variable	Coefficient Estimates [p-value]							
Factor: Likability	**-1.285**	**-1.258**	**-2.459**	**-2.234**	-0.850	**-1.253**	**-1.885**	**-1.799**
	[0.04]	[0.00]	[0.04]	[0.00]	[0.13]	[0.00]	[0.08]	[0.00]
Factor: Difficulty	**1.831**	**1.261**	**3.145**	**2.057**	**1.323**	**1.213**	**1.959**	**1.316**
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.02]	[0.01]
Intermediate Microeconomics Grade	**0.391**	**0.385**	**0.401**	**0.387**	**0.391**	**0.386**	**0.401**	**0.387**
	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]	[0.00]
Gender	0.060	0.026	0.156	0.040	0.058	0.021	0.147	0.040
	[0.55]	[0.69]	[0.24]	[0.65]	[0.56]	[0.75]	[0.27]	[0.65]
Underrepresented Minority	-0.020	**-0.179**	**-0.344**	**-0.354**	-0.018	**-0.177**	**-0.345**	**-0.358**
	[0.89]	[0.07]	[0.07]	[0.01]	[0.90]	[0.07]	[0.07]	[0.01]
Low Income	**-0.290**	**-0.168**	**-0.299**	**-0.241**	**-0.299**	**-0.170**	**-0.315**	**-0.248**
	[0.02]	[0.03]	[0.10]	[0.02]	[0.02]	[0.03]	[0.08]	[0.02]
Language	0.065	**-0.094**	0.079	-0.038	0.066	**-0.092**	0.079	-0.039
	[0.32]	[0.02]	[0.36]	[0.48]	[0.31]	[0.02]	[0.36]	[0.46]

First Generation	-0.091	**-0.172**	-0.190	**-0.216**	-0.091	**-0.172**	-0.190	**-0.215**
	[0.45]	[0.02]	[0.25]	[0.02]	[0.45]	[0.02]	[0.25]	[0.02]
SAT Math	**0.001**	---	**0.003**	---	**0.001**	---	**0.003**	---
	[0.09]		[0.00]		[0.10]		[0.01]	
SAT Writing	**0.002**	---	0.000	---	**0.002**	---	0.000	---
	[0.02]		[0.70]		[0.02]		[0.74]	
Repeat	**-0.620**	**-0.967**	-0.119	**-0.938**	**-0.623**	**-0.972**	-0.110	**-0.936**
	[0.01]	[0.00]	[0.66]	[0.00]	[0.01]	[0.00]	[0.69]	[0.00]
Summer Course	**-1.534**	**-0.981**	---	---	**-1.373**	**-0.850**	---	---
	[0.00]	[0.01]			[0.01]	[0.02]		
Instructor Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Evaluation Questions Included	10	10	12	12	10	10	12	12
Pseudo R^2	0.096	0.083	0.101	0.092	0.095	0.083	0.101	0.091
Observations	1480	3157	843	1803	1480	3157	843	1803

The results from Table 5A are relatively robust across the various estimated models. Coefficients in bold indicate variables that are statistically significant at a 10% level. Likability negatively affects the probability of earning a higher grade in the follow-on elective course while difficulty positively affects that probability. From Table 5A we focus our attention on the specifications that use the factors based loadings and include the SAT scores, one with the 10 questions, and one with twelve questions (the first and third column of coefficients). These are the models that have the highest pseudo R^2 (marginally). We then calculate the marginal effects from these two models. As we are interested in the impact of the identified factors, we present the marginal effects for those two factors for every level of the advanced grade earned (i.e. grades A to F). The results are in Table 5B.

Table-5B. Marginal Effects

	Using Ten Questions		**Using Twelve Questions**	
Grade	Likability	Difficulty	Likability	Difficulty
A	-0.158	0.226	-0.319	0.408
	[0.04]	[0.00]	[0.04]	[0.01]
A-	-0.097	0.139	-0.193	0.246
	[0.04]	[0.00]	[0.04]	[0.01]
B+	-0.055	0.079	-0.096	0.123
	[0.05]	[0.00]	[0.05]	[0.01]
B	0.010	-0.014	0.068	-0.087
	[0.26]	[0.20]	[0.09]	[0.04]
B-	0.061	-0.087	0.130	-0.166
	[0.04]	[0.00]	[0.04]	[0.01]
C+	0.070	-0.099	0.140	-0.179
	[0.04]	[0.00]	[0.04]	[0.01]
C	0.091	-0.130	0.163	-0.209
	[0.04]	[0.00]	[0.04]	[0.01]
C-	0.031	-0.044	0.039	-0.050
	[0.05]	[0.00]	[0.05]	[0.01]
D+	0.014	-0.020	0.014	-0.018
	[0.06]	[0.00]	[0.09]	[0.04]
D	0.016	-0.023	0.025	-0.032
	[0.06]	[0.00]	[0.06]	[0.02]
D-	0.005	-0.008	0.008	-0.011
	[0.08]	[0.01]	[0.12]	[0.07]
F	0.013	-0.019	0.021	-0.027
	[0.05]	[0.00]	[0.05]	[0.01]

Table 5B highlights the effects of a one unit change in the likability and difficulty factors on the probability of a student earning a particular grade in their advanced elective courses. The p-value for each estimated marginal effect based on the z-scores is also provided. Nearly every marginal effect is significant at a 10% level. We can see that grades above "B" are when difficulty of the intermediate microeconomics instructor begins to positively contribute to student's performance in their future courses while likability begins to negatively affect it. For example, from the table, with the factors using ten questions from the evaluation form , a one unit increase in the likability factor decreases the probability a student earns a B+ in their follow-on course by 5.5 percentage points while a one-unit increase in difficulty increases the probability of a B+ by 7.9 percentage points. Of course it is difficult to interpret the meaning of a one-unit increase in these factors due to the inherent nature of artificially constructed indices that

we believe are proxies for latent variables. For example, Table 6 shows the mean, median, and standard deviation of each of the factors.

Table-6. Factor Loadings Statistics

	Using Ten Questions		Using Twelve Questions	
Statistic	Likability	Difficulty	Likability	Difficulty
Mean	4.16	4.24	4.09	4.29
Median	4.19	4.28	4.12	4.28
Standard Deviation	0.32	0.24	0.36	0.17

As we can see from Table 6, the factor loadings are slightly larger than 4 on average with a standard deviation ranging from 0.17 to 0.36. Thus a one-unit change in the factors used in Table 5B appears to be substantial. However, again, it is hard to know the precise interpretation of the factors and thus hard to understand the magnitude. Still, Table 5B illustrates a clear connection between the general student evaluation of the instructors in intermediate microeconomics and the performance of individual intermediate microeconomics students in their future related courses.

6. Conclusion

How a student learns and retains information along with the combination of internal and external factors influencing that is likely highly complex and perhaps fluid. In this study we attempt to isolate the role that a prior instructor of a core course plays in that learning in immediately subsequent courses. Our findings indicate that an instructor's likability appears to hinder future student learning while difficulty enhances it. We cannot say what the precise mechanisms are that lead to that result. Delving into that process would be fruitful avenue for future research.

Of course an easy criticism of this work is that the identified factors of likability and difficulty are not really those traits at all. Rather they could be something else. That is true. However, as we are instructors that interact routinely with other instructors we often hear that the overall teaching rating for an instructor is nothing more than an "applause meter", a measure of how enjoyable the course was or likable the instructor was during that quarter. In our specifications, the likable factor includes that overall rating every time. The direct implication is that the item that is often a heavily weighted data point in the evaluation of an instructor's teaching for merit and promotion is not a good proxy for that. Other information contained in the student evaluations may be better used in that regard. Further research to confirm this finding would be most beneficial, especially in other contexts and courses beyond intermediate microeconomics.

Finally, the strongest loadings in the difficulty factor are the statements where students believe the exams were fair, timely grades, and the workload of the course appropriate to the number of units. The instructional value of the assignments in the course also loads on the difficulty factor. Perhaps the old adage that it is best to be tough, but fair, is what is being found in this study. Instructors that are difficult are so because they are tough but fair and possibly that combination creates a longer-lasting retention of the information from the course for students that those with instructors that are likable but not particularly memorable. While an instructor can be both difficult and likable, finding the right level of difficulty that increases retention of information for students may be the critical element to identify and cultivate.

References

Aaronson, D., Barrow, L. and Sander, W. (2007). Teachers and student achievement in the Chicago public schools. *Journal of Labor Economics*, 25(1): 95-135.

Anne and Boring (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145(1): 27-41.

Brian, A., Jacob, Lars and Lefgren (2004). The impact of teacher training on student achievement: Quasi-experimental evidence from school re- form efforts in Chicago. *Journal of Human Resources*, 39(1): 50–79.

Bruce, A., Weinberg, Masanori, Hashimoto., Belton, M. and Fleisher (2010). Evaluating teaching in higher education. *The Journal of Economic Education*, 40(3): 227-61.

Charles, T., Clotfelter, Helen, F. L., Jacob, L. and Vigdor (2006). Teacher sorting, teacher shopping, and the assessment of teacher effectiveness. *Journal of Human Resources*, 41(4): 778-820

Florian, Hoffmann, Philip and Oreopoulos (2009). Professor qualities and student achievement. *The Review of Economics and Statistics*, 91(1): 83-92.

Jonah, E. and Rockoff, 2004. "The impact of individual teachers on student achievement: Evidence from panel data, the American economic review." In *Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association San Diego, CA.* pp. 247-52.

Kane, T. J., Rockoff, J. E. and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, 27(6): 615-31.

Kane, T. J., Taylor, E. S., Tyler, J. H. and Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, 46(3): 587-613.

Lillian, M., Adam, D. and Andrea, H. (2014). What's in a name: Exposing gender bias in student ratings of teaching. *Innovation Higher Education*, 40(4): 291-303.

Rivkin, S. G., Hanushek, E. A. and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2): 417-58.

Scott, E., Carrell, James, E. and West (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3): 409-32.

Thomas (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources Summer*, 46(3): 587-613.