Does Class Size Matter in the University Setting?

Ethan Ake-Little[1], Nathaniel von der Embse[2], and Dana Dawson[1]

University class size is a frequently debated topic among stakeholders given its relation to student achievement, teaching and learning, program evaluation, and education economics. However, the extant literature in both K–12 and higher education contexts regarding class size is equivocal, with some citing evidence of an adverse effect on student achievement for larger class sizes and others suggesting minimal effect. This study aims to explore the relationship between class size and student achievement, as measured by course grades, in the core undergraduate program at Temple University, a large, state-related university in Philadelphia, Pennsylvania. A cross-classified multilevel model was employed consisting of 14 variables—6 student level and 8 class-level—and drawing from a robust sample size of 172,516 grades awarded to 32,766 students in 8,049 classes offered across 14 terms. Results suggest that, after controlling for instructor experience, the effect of class size is not uniform and is, in fact, quite variable when accounting for student race, gender, and academic discipline within the same model. We discuss the possible reasons for these variable results with implications for program policy and classroom practice. Finally, we discuss the limitations of the present study and how future research might resolve those limitations.

Keywords: class size; multilevel modeling; program evaluation; student achievement; university teaching and learning

Class size is an issue that continues to stimulate debate among educators, researchers, administrators, and policymakers because of its link to student achievement, the quality of teaching and learning, and program evaluation. Furthermore, class size has also received considerable attention from stakeholders over the past quarter century given its economic implications, particularly for institutions of higher education (IHEs) (Krueger, 2003). In a 2014 report published by *The Chronicle of Higher Education* that examined state aid to public colleges between 1987 and 2012, 164 out of 169 research universities reported a decline in state funding within that period, with nearly one in three reporting a decrease of 25% or more. Consequently, IHEs may seek to increase class size as a means of reducing human resource costs.

Most studies related to class size, whether K–12 or higher education, tend to be highly quantitative (e.g., Diette & Raghav, 2015; Hanushek, 1986; Hoxby, 2000; Kokkelenberg et al., 2008). However, their conclusions regarding the influence of class size on student achievement are equivocal at best. In the K–12 literature, some research has suggested that class size does influence student achievement at various points in the K–12 experience (Borland et al., 2005). However, other research has suggested that teacher efficacy has a more salient influence on student achievement (Hoxby, 2000; Rockoff, 2004). When studying IHEs, researchers have similar disagreements as to the influence of class size on student achievement and its implications (Keil & Partell, 1997; Kokkelenberg et al., 2008; Terenzini & Pascarella, 1991; Williams et al., 1985). Methodologically, these class size studies employ a diverse array of models that include student achievement controls and classroom environment variables; however, very few control for instructor experience or have attempted to parse the effect of class size by simultaneously including race, gender, and academic discipline within the same model.

Given the methodological limitations of previous higher education class size studies, there is a need for a more robust quantitative analysis that incorporates a broader range of student and class-level variables, including those that control for instructor experience and student experience. The goal of this study is to present such a comprehensive model aimed at addressing this gap in the literature.

[1]Temple University, Philadelphia, PA
[2]University of South Florida, Tampa, FL

Educational Researcher, Vol. XX No. X, pp. 1–11
DOI: 10.3102/0013189X20933836

Literature Review

Class Size in the K–12 Context

Class size has been examined extensively using experimental and quasi-experimental designs, both of which seek to determine a causal relationship between resource allocation and student outcomes. Hanushek (1986) found that "there is not a strong or consistent relationship between student performance and school resources" such as class size. However, Krueger (2003) called into question this analysis, arguing that Hanushek based his conclusions on 277 estimates drawn from 59 studies with all studies weighted equally. Because some studies account for a proportionally higher number of calculations, smaller subsets of data have an outsized influence. When Krueger reanalyzed the data with new weighting, results indicated a positive relationship between school resources and student performance.

The Tennessee Student Teacher Achievement Ratio (STAR) Experiment was a randomized trial in which over 11,500 students in K–3 grades and 1,300 teachers in 79 Tennessee elementary schools were randomly assigned to small ($\leq$20 students) or regular-sized classes ($>$20 students) from 1985 to 1989. Students' achievement in math and reading standardized tests in the STAR Experiment improved by about 0.15 to 0.20 standard deviation in a small class as opposed to a regular-sized class (Word et al., 1990). When results were disaggregated by race, Black students as well as students of low socioeconomic background (as indicated by participation in the Free or Reduced Price Lunch program) experienced more significant gains from being assigned to a small class (Krueger & Whitmore, 2002).

More recently, researchers have attempted to move away from an approach that seeks a single answer regarding class size and toward a more contextualized approach by incorporating grade level or program type. Borland et al. (2005) argue that the literature surrounding class size has been inconclusive because of erroneous assumptions in the data, including errors associated with using a student-teacher ratio and failure to control for student ability. When the authors controlled for these assumptions, they found that the relationship between class size and student achievement is both nonlinear and nonmonotonic, suggesting the impact of class size on student achievement can vary significantly from grade to grade and subject to subject. In a study of class size and student achievement between the fourth and sixth grades, Hoxby (2000) concluded that class size makes no difference when a variation in the student population triggers an additional section (i.e., if the 26th student added to a class of 25 results in two sections of 13). Moreover, her findings suggest no class-size effects at schools that serve disproportionately large shares of disadvantaged or minority students.

The disparities between studies have led researchers to believe that instruction and pedagogy may influence the relationship between class size and student outcomes. Smaller class sizes provide the most benefit for low-achieving students only when they receive individual attention, which, in turn, facilitates their learning (Blatchford et al., 2011).

Class Size in the Higher Education Context

The complex models conceived in the K–12 context are arguably more comprehensive than models designed to study class size in the university setting. Smith and Glass (1980) incorporated both teacher and student perceptions related to instruction in large classes, and the STAR Experiment included student achievement and instructor demographic data. Additionally, the STAR Experiment included universalized measures such as standardized test scores that made global comparisons possible. Researching the issue of class size in higher education, however, is more complicated, and the models developed to study this phenomenon lack the level of sophistication previously seen in the K–12 literature. Kokkelenberg et al. (2008) modeled how class size influenced the end-of-course grade with and without fixed effects using over 760,000 undergraduate student grades from a northeastern public university. They concluded that class size has a negative relationship with student grades. Grade point averages declined significantly for class sizes up to 20 students and more gradually for larger classes. Although their model did include variables related to student academic achievement, it did not account for either student demographics, instructor experience, or the classroom experience. Similarly, Diette and Raghav (2015) conducted an analysis using data from a selective liberal arts college and concluded that student grades decrease as class size increases, particularly for more vulnerable students such as freshman students or those with low SAT scores. Like Kokkelenberg et al. (2008), their model draws from a large sample size (n = 134,195) but only tangentially addresses the issue of instructor experience by adding variance terms for both department and faculty.

As with K–12 studies, there have been arguments made against the relationship between class size and student achievement. Some research concludes that class size has little to no influence on student achievement, arguing that instruction steeped in "lower-level" educational outcomes, such as recalling of facts, would be mostly unaffected by increasing class size (Terenzini & Pascarella, 1991; Williams et al., 1985). McKeachie (1980) and Iran-Nejad et al. (1990) concede that this is likely the case for courses that rely on traditional lecturing but argue that smaller classes are the best environment for promoting critical thinking and advanced problem-solving. It is important to note that none of these studies examined the effect of class size on specific domains of coursework or if class size had an influence on specific student demographics.

Outcomes in Higher Education

Whereas multiple outcomes exist in the K–12 context, both standardized and unstandardized, making it possible to explore the issue of class size through several avenues, the outcome variable presents a unique challenge in higher education. The de facto variable in most higher education class size studies is student grades, in part because it is universal in terms of measurement and effect. Although student course evaluations are also a popular outcome and can help explain student perceptions between class size and learning, they are limited in their generalizability because of their design and measurement variability by institution. In a

study of student evaluations of economics courses at the University of California, Santa Barbara, from 1997 to 2004, Bedard and Kuhn (2008) reported a large, highly significant, and nonlinear negative impact of class size on student evaluations of instructor teaching. Similar results surface across economic disciplines and types of universities (Monks & Schmidt, 2011; Walia, 2008). Although all three studies probed student perceptions of the instructor, none of them controlled for instructor experience.

The inclusion of student evaluation data is not without controversy. Some argue that students may be motivated to complete evaluations with a primary concern towards their grade in the course rather than critically evaluating the teaching and learning dynamic (Watchtel, 1998). Others have noted that students tend not to put much emphasis on class size when choosing a university; however, once they begin taking classes, they develop an aversion to large classes (Drewes & Michael, 2006). Likewise, Carbone and Greenberg (1998) found that only a quarter of U.S. college students surveyed felt that class size did not impact their learning.

Although many class size studies in higher education control for student ability, generally through some combination of GPA and standardized test scores, few appear to control for instructor experience. To address this issue, Arias and Walker (2004) designed an experimental study with the same instructor, course, and content using two large classes (90 students) and two small classes (25 students). When they compared total exam scores, they found a statistically significant negative relationship between class size and student performance, with students in small classes performing only 3% higher on exams. When student-level data from eight cohorts of first-year students at Northwestern University was examined to investigate the relative effects of tenure-track/tenured versus contingent faculty on student learning, Figlio et al. (2015) concluded that students learn relatively more from contingent faculty in their first-term courses.

Purpose of the Present Investigation

This study aims to understand the possible effects of increasing class size on student achievement in the context of Temple University's General Education (GenEd) program. The program is the required curriculum for all undergraduate students, consists of 11 interdisciplinary courses in 10 areas generally taken in the first 2 years, and is designed to expose students to a variety of disciplinary knowledge and foundational skills such as reading and writing, numerical literacy, and scientific reasoning. In addition to examining the relationship between class size, student demographics, and student achievement, the study also attempts to explain how the effect may vary across academic domains. Therefore, we seek to answer the following research questions in this study:

What, if any, is the nature of the relationship between student race, gender, and varying class sizes on student achievement in

(1) Social science GenEd courses?
(2) STEM (science, technology, education, and mathematics) GenEd courses?
(3) Arts & humanities GenEd courses?

Methods

Participants

The population consists of 287,243 grades (observations) earned by 54,319 students in 10,152 GenEd course sections over 14 semesters from Fall 2011 through Spring 2016 terms. The population includes only non-Honors, non-online, and single instructor courses. These criteria were used because (a) Honors level courses and courses with multiple instructors (e.g., recitations, labs, studio) have restrictions on class size, thereby unduly influencing the sample; and (2) online courses may be virtual, hybrid, or asynchronous, and therefore have online in-person meetings, some in-person meetings, or no in-person meetings at all.

Three sources of data were collated to produce the population data set. First, the Banner database provided student-level data that contained the student's ID, race, gender, and GenEd course information including class size and final course grade over the study period. We converted letter grades into their 4.0 equivalents (e.g. A– = 3.67), and all cases with a nonstandard letter grade (e.g., I = Incomplete, M = Medical) were removed from the dataset. Next, these data were merged with admissions data that included the student's high school GPA and SAT Math and Verbal scores. For students with ACT scores, we converted them into their SAT equivalents as per the official Educational Testing Service (ETS) concordance table (Dorans, 1999). Third, this dataset was combined with student course evaluation data obtained from Institutional Research and Assessment (IRA), which included three instructor-experience-related variables (instructor rank, the number of years of teaching experience, and the number of times the instructor taught the course) and four student evaluations measures (level of interest, expected grade in course, hours spent per week preparing for the course, and overall student perception of their preparation), the latter of which were aggregated at the course level. Finally, for each observation, we classified the 10 GenEd areas into three domains: Social Sciences, STEM, and Arts & Humanities.

Variables and Measures

The model consists of 14 variables—6 student level and 8 class level—with the outcome being the course grade (on a 4.00 scale), which was treated continuously. The remaining five student-level variables in this model are high school GPA (measured on a 5.00 scale, which is the result of an admissions formula that weights for school and coursework quality), SAT Math and Verbal (measured on a 200–800 point scale), student race, and gender. Multiple self-reported student racial categorizations in the dataset were simplified into three categories based on their representation in higher education: Underrepresented Groups (African Americans, Hispanics, and American Indians), Overrepresented Groups (Whites, Asian/Pacific Islanders), and Other Groups (Multiracial and Unknown).

Similarly, at the classroom level, instructor experience variables were also reorganized. Instructor rank comprises three categories: Contingent Faculty (Graduate Assistant, Adjunct Instructor and Non-Tenure-Track Instructor), Tenure Track Faculty (Tenure-Track Instructors, Tenured Instructors), and Other Faculty (Visiting/Volunteer Faculty, Unknown Faculty). Instructor

Table 1
Descriptives for All Continuous Variables Used in Model (*n* = 172,516)

	Mean	*SD*	Measurement Scale
(Raw) course grade	3.17	0.93	4.0 scale
High school GPA (weighted)	3.35	0.49	5.0 scale
SAT Math score	550.53	75.71	200–800 scale
SAT Verbal score	540.16	76.57	200–800 scale
(Class) interest in course[a]	1.92	0.29	1 = low interest, 2 = medium interest, 3 = high interest
(Class) expected grade in course[a]	3.41	0.29	4.0 scale
(Class) hours per week spent preparing for class[a]	3.43	0.68	1 = Less than 1 hour 2 = 1 to 2 hours 3 = 2 to 3 hours 4 = 3 to 4 hours 5 = 4 to 6 hours 6 = 6 to 8 hours 7 = 8 or more hours
(Class) came well prepared to class[a]	4.14	0.27	5.0 scale

[a]Originally student-level data but was provided as class level aggregates. Thus, categorical items were converted to continuous measures.

experience also comprises three categories: New Instructor (0–5 Years' Experience); Experienced Instructor (6–9 Years' Experience), and Master Instructor (10+ Years' Experience); along with times taught, which was restructured into four categories: Low Repeat (1–4x), Moderate Repeat (5–10x), High Repeat (10+x), and Other (Declined to Answer). Since IRA provided the course evaluation variables as class aggregates, they were transformed from categorical to continuous variables on a 5.0 scale. Finally, because class size is a nonlinear continuous variable, it too was recategorized based on percentile rank in the population:

- Small Classes: Up to 25th Percentile (≤25 Students)
- Medium Classes: 26th–50th Percentile (26–30 Students)
- Large Classes: 51st–70th Percentile (31–40 Students)
- Extra Large Classes: 71st–80th Percentile (41–-60 Students)
- Oversized Classes: ≥80th Percentile (≥61 Students)

Tables 1 and 2 provide descriptive statistics for all continuous and descriptive variables, respectively, and Table 3 provides descriptive statistics for letter grades awarded by class size groupings and academic discipline.

Data Analytic Plan

We employed a random intercept only cross-classified multilevel model since this study examines observations (course grades) clustered by students, class sections, and even term to account for the longitudinal nature of the data. Although we predict that the effect of class size varies *within* classes (particularly in terms of student demographics), we assume that the effect is constant *across* classes, thus favoring a random intercept only model. Moreover, we elected to use the raw outcome variable since, despite having a skewness value of –1.57, it was within the acceptable range of ±2 (Bryne, 2016; Kline, 2011). Given the clustering of these data, a cross-classified multilevel model is a superior alternative to traditional regression techniques, which, in this case, would provide downwardly biased standard errors (thereby increasing the probability of committing Type I error) as well as less efficient estimates of regression coefficients (Snijders & Bosker, 1999). The cross-classified multilevel model was created using R's lme4 package. (See the supplementary online appendix, available on the journal's website, for R code.) R's lme4 function performs only listwise analyses, resulting in a final dataset of 172,516 grades (about 60% of the population) earned by 32,766 students across 8,049 sections offered in the 14-term period. We opted to use full maximum likelihood estimation to test the model fit between the fixed and random effects. Although full maximum likelihood underestimates the variance parameters because it assumes that the fixed effect parameters are known with certainty when estimating these variance parameters, the large sample size is likely to mitigate these concerns (Snijders & Bosker, 1999).

Three of the six student-level predictors (high school GPA, SAT Math, SAT Verbal) and one class-level variable (expected grade in the course) were *z*-standardized and group-centered for use in this model. Group centering allows for the within-group variable to be captured, thus permitting more accurate estimates with smaller standard errors when attempting to ascertain the relationship between a class-level variable (class size) and a student-level variable (course grade) (Snijders & Bosker, 1999). Additionally, a subsidiary analysis akin to a propensity analysis revealed that there was no selection bias on the part of students electing to enroll in each class size grouping based on their prior ability. Finally, class-level variables—class size, course evaluation measures, and instructor experience variables—are idiosyncratic to a given class; semester changes in these measures are captured in the dataset. A summary equation of the model is as follows:

$$Y_{ijk} = \beta_0 + \beta_1 * HSGPA_{ijk} + \beta_2 * SATMATH_{ijk} + \beta_3 * SATVERB_{ijk} + \beta_4 * STUDGEND_{ijk} + \beta_5 * STUDRACE_{ijk} + \mu_{0i} + \mu_{0j} + \mu_{0k} + \varepsilon_{ij}$$

Table 2
Descriptives for All Categorical Variables Used in Model (Dummy Coded)

	n	Mean
Student gender		
Male	81,628	0.473
Female	90,888	0.527
Student race		
Overrepresented groups	127,767	0.741
Underrepresented groups	31,330	0.182
Other/unknown	13,419	0.078
GenEd course disciplines		
Social science courses	74,694	0.433
STEM courses	25,185	0.146
Arts & humanities courses	72,637	0.421
Class size groupings		
Small classes (up to 25th percentile: 0–25 students)	41,438	0.240
Medium size classes (26th–50th percentile: 26–30 students)	43,068	0.250
Large size classes (51st–70th percentile: 31–40 students)	34,762	0.202
Extra large size classes (71st–80th percentile: 41–60 students)	20,062	0.116
Oversize classes (>80th percentile: >60 students)	33,186	0.192
Instructor rank		
Contingent instructor	130,707	0.758
Tenure-track/tenured instructor	19,890	0.115
Other/declined to answer	21,919	0.127
Instructor years of teaching experience		
New teacher (≤5 years' experience)	52,407	0.303
Experienced teacher (6–9 years' experience)	20,418	0.118
Master teacher (>10 years' experience)	29,396	0.170
Declined to answer	70,925	0.410
Number of times instructor taught course		
Low repeat (<5x)	70,395	0.408
Medium repeat (5–10x)	39,844	0.231
High repeat (>10 times)	37,590	0.218
Declined to answer	24,687	0.143

where Y_{ijk} is the course grade for the *i*th student in the *j*th classroom during the *kth* term; β_0 is the random intercept that includes all class-level variables, as well as all interaction terms between student race, gender, GenEd course domain, and various class size groupings; β_1 through β_5 are the fixed effects for the student-level variables (absent the outcome); and the terms $\mu_{0i} + \mu_{0j} + \mu_{0k}$ are the variance terms associated the student, class (section), and term effects, respectively, along with ε_{ijk} as the error term associated with each student.

Results

Three models were created for this analysis; the first two models are identical, with the exception that the first does *not* include the four class-level course evaluation metrics, whereas the second does include them. Both are presented alongside the third model, which compounds the superior alternative with interaction effects. The rationale for this setup was to compare if, and to what extent, these four variables were endogenous to the outcome; the results suggest that, given the highly similarly qualitative properties of these results (similar vectors and significance levels), these variables are likely not endogenous to the outcome. Since the model with the evaluation metrics is a better fit, as evidenced by Akaike information criterion (AIC) and Bayesian information criterion (BIC) values, the second model was used to build the final model with interaction effects (Ebbes et al., 2004; Podsakoff et al., 2003).

The reference category for the third model is a (a) White or Asian/Pacific-Islander (b) male student enrolled in (c) a social science GenEd course (d) with a small class size (e) taught by a contingent faculty member (f) who has less than or equal to 5 years of experience (g) and has taught the course less than five times. Table 4 outlines the main effects for all control variables used, and Table 5 parses the effect of class size for all combinations of student race, gender, academic discipline, and class size groupings. The average course grade is a high B ($M = 3.17$, $SD = 0.93$).

Interaction Effect for Social Science GenEds

As expected, the effect of class size on course grade in social science GenEds varies considerably by student demographic. Underrepresented males perform worse in the small classes, $\beta = -0.148$, $t(515) = -3.992$, $p < .001$, compared to the larger ones; but this same class size grouping has no statistically significant effect on underrepresented females, who experience an increase in course grade in oversize classes, $\beta = 0.122$, $t(2,057) = 2.375$, $p < .05$. Overrepresented groups have a converse pattern, with males performing worse in oversize classes, $\beta = -0.079$, $t(7,189) = -3.202$, $p < .01$; and females performing better in small, $\beta = 0.184$, $t(2,885) = 8.854$, $p < .001$, and medium classes (26–30 students), $\beta = 0.062$, $t(3,516) = 2.380$, $p < .05$.

Interaction Effect for STEM GenEds

The effect of class size on course grade in STEM GenEds shares similarities with the effect in social science GenEds. Here, underrepresented males also perform worse in the small classes, $\beta = -0.179$, $t(195) = -2.772$, $p < .01$, compared to progressively larger classes, but this increase appears to have no statistically significant effect on underrepresented females. For overrepresented groups, again there is a converse pattern, with White and Asian/Pacific-Islander males performing better in small classes, $\beta = 0.114$, $t(961) = 3.534$, $p < .001$, but performing worse in extra-large classes, $\beta = -0.174$, $t(1,224) = -3.667$, $p < .001$. White and Asian/Pacific-Islander females do not appear to experience a detrimental effect until placed in oversized classes, $\beta = -0.140$, $t(2,459) = -3.271$, $p < .01$.

Interaction Effect for Arts & Humanities GenEds

The effect of class size on course grade in arts & humanities GenEds has a pattern distinct from the other two domains.

Table 3
Cross-Tabulations by General Education Course Type, Class Size Grouping, and Range of Grade Awarded (n = 172.516)

	Social Science Courses									
	A Range Grade		B Range Grade		C Range Grade		D Range Grade		F Grade	
	n	Mean	*n*	Mean	*n*	Mean	*n*	Mean	*n*	Mean
Class size groupings										
Up to 20th percentile (≤25 students)	3,906	0.112	2,480	0.092	837	0.103	182	0.107	424	0.143
21st–50th percentile (26–30 students)	4,880	0.140	2,781	0.103	914	0.112	205	0.121	389	0.132
51st–70th percentile (31–40 students)	12,286	0.351	9,453	0.351	2,701	0.331	544	0.321	969	0.328
71st–80th percentile (41–60 students)	6,408	0.183	4,850	0.180	1,281	0.157	240	0.142	424	0.143
>80th percentile (>60 students)	7,496	0.214	7,342	0.273	2,428	0.298	525	0.310	749	0.253
Total (across all class size groupings)	34,976	1.000	26,906	1.000	8,161	1.000	1,696	1.000	2,955	1.000
	STEM Sciences Courses									
	A Range Grade		B Range Grade		C Range Grade		D Range Grade		F Grade	
	n	Mean	*n*	Mean	*n*	Mean	*n*	Mean	*n*	Mean
Class size groupings										
Up to 20th percentile (≤25 students)	1,362	0.123	917	0.101	374	0.102	54	0.107	150	0.178
21st–50th percentile (26–30 students)	2,974	0.269	1,709	0.188	540	0.147	79	0.156	136	0.162
51st–70th percentile (31–40 students)	2,918	0.264	2,153	0.237	749	0.204	131	0.259	157	0.187
71st–80th percentile (41–60 students)	1,687	0.152	1,385	0.152	403	0.110	48	0.095	102	0.121
>80th percentile (>60 students)	2,130	0.192	2,936	0.323	1,602	0.437	193	0.382	296	0.352
Total (across all class size groupings)	11,071	1.000	9,100	1.000	3,668	1.000	505	1.000	841	1.000
	Arts & Humanities Courses									
	A Range Grade		B Range Grade		C Range Grade		D Range Grade		F Grade	
	n	Mean	*n*	Mean	*n*	Mean	*n*	Mean	*n*	Mean
Class size groupings										
Up to 20th percentile (≤25 students)	12,918	0.388	12,622	0.447	3,558	0.468	549	0.447	1,105	0.479
21st–50th percentile (26–30 students)	13,544	0.407	11,347	0.402	2,515	0.331	325	0.265	730	0.316
51st–70th percentile (31–40 students)	1,471	0.044	811	0.029	244	0.032	52	0.042	123	0.053
71st–80th percentile (41–60 students)	1,807	0.054	960	0.034	307	0.040	58	0.047	102	0.044
>80th percentile (>60 students)	3,532	0.106	2,488	0.088	978	0.129	243	0.198	248	0.107
Total (across all class size groupings)	33,272	1.000	28,228	1.000	7,602	1.000	1,227	1.000	2,308	1.000
Total (across all disciplines)	79,319	0.460	64,234	0.372	19,431	0.113	3,428	0.020	6,104	0.035

Interestingly, neither underrepresented males nor females appear to be affected by increases in class size. For overrepresented groups, males perform better in medium classes, $\beta = 0.062$, $t(10{,}725) = 2.268$, $p < .05$; whereas females perform worse in medium classes, $\beta = -0.083$, $t(10{,}692) = -2.386$, $p < .01$. Moreover, these groups perform better in extra-large classes, $\beta = 0.096$, $t(1{,}231) = 2.461$, $p < .05$.

Model Fit Statistics

The student, class, and term intraclass correlations (ICCs) for the third model are .409, .077, and .008, respectively, suggesting that the 172,516 observations clustered within 32,766 students, 8,049 class sections, and 14 terms have an altogether weak correlation with one another. Moreover, when comparing the AIC and BIC best fit indices, the third model, which contains all four course-evaluation measures along with several interaction effects, was superior to the first model, which did not contain either the course evaluation measures or any interaction effects.

Discussion

Although research on the influence of class size has been somewhat inconclusive, our findings suggest that this ambiguity may

Table 4
Main Effects of Model Variables on Transformed GenEd Course Grade (n = 172,516)

	Outcome: Course Grade (4.00 scale)		
	Model Without Course Evaluation Responses	Model With Course Evaluation Responses	Model With Course Evaluation Responses + Interactions
Fixed effects			
Intercept	3.178 (0.022)***	3.142 (0.069)***	3.130 (0.070)***
Student level—Prior academic ability			
High school GPA (weighted)[a]	.188 (.004)***	.186 (.004)***	.186 (.004)***
SAT Math score[a]	.014 (.004)***	.013 (.004)**	.014 (.004)**
SAT Verbal score[a]	.069 (.004)***	.070 (.004)***	.070 (.004)***
Student level—Demographics			
Female	.178 (.008)***	.176 (.008)***	.184 (.021)***
Underrepresented minorities	−.199 (.010)***	−.196 (.010)***	−.148 (.037)***
Class level—Class size categories			
Medium (26th–50th percentile: 26–30 students)	.068 (.010)***	.025 (.009)**	−.020 (.024)
Large (51st–70th percentile: 31–40 students)	.040 (.013)**	−.033 (.011)**	−.025 (.020)
Extra large (71st–80th percentile: 41–60 students)	.043 (.016)**	−.017 (.014)*	.009 (.023)
Oversize (>80th percentile: >60 students)	−.066 (.018)***	−.085 (.015)***	−.079 (.025)**
Class level—Course discipline			
STEM GenEd courses	−.043 (.013)**	−.037 (.012)**	.114 (.032)***
Arts & humanities GenEd courses	−.032 (.011)**	.009 (.009)	.003 (.019)
Class level—Student interest & preparation			
Level of interest	—	.023 (.014)	.028 (.014)*
Expected grade	—	.609 (.014)***	.609 (.014)***
Hours spent per week preparing for class	—	−.063 (.006)***	−.063 (.006)***
Came well prepared to class	—	.031 (.013)*	.029 (.013)*
Class level—Instructor experience			
Tenure track/tenured instructor	−.133 (.016)***	−.055 (.013)***	−.056 (.013)***
Teaching exp: Experienced (6–9 yrs.)	−.042 (.015)**	−.030 (.012)*	−.032 (.012)**
Teaching exp: Master (≥10 yrs.)	−.075 (.014)***	−.044 (.012)***	−.045 (.012)***
Times taught: Medium repeat (5–10x)	.015 (.011)	.007 (.009)	.008 (.009)
Times taught: High repeat (>10x)	.003 (.012)	.009 (.010)	.009 (.011)
Interaction effects[b]	—	—	See Table 5
Random effects			
Student (σ^2)	.322	.322	.322
Class (σ^2)	.098	.061	.061
Course term (σ^2)	.004	.006	.006
Residual (σ^2)	.398	.399	.399
Students	32,766	32,766	32,766
Classes (sections)	8,049	8,049	8,049
Terms	14	14	14
Log likelihood	−196,667.100	−195,379.700	−195,173.000
Akaike information criterion (AIC)	393,384.200	390,817.400	390,563.900
Bayesian information criterion (BIC)	393,635.700	391,909.100	391,660.300
Student intraclass correlation (ICC)	.392	.409	.409
Classroom ICC	.119	.077	.077
Term ICC	.005	.008	.008

Note. Declined to answer/other categories have been omitted for simplicity.

[a]z-standardized (group centered).

[b]For interactions effects of gender, race, and course discipline on class size, please see Table 5.

*$p < .05$. **$p < .01$. ***$p < .001$.

be because the effect of class size is far more nuanced than historically discussed. Accordingly, the profile of our university—an R1 state-related institution that is amongst the top 50 in terms of student enrollment and has a combined average SAT Math and Verbal score of 1,090 (55th percentile)—makes it possible to apply our findings and recommendations to a myriad of public institutions. Finally, given the large sample size of this analysis (n = 172,516) and the robust number of observations within each interaction effect (Table 5), even the absence of statistically significant findings in some interaction groupings is likewise telling (Kline, 2011).

In terms of student race and gender, the findings for underrepresented groups contrast with previous research, which has found that smaller class sizes correlate with improved academic outcomes. When considering small class size outcomes, male students who are African American, Hispanic, and American Indian perform worse in social science GenEds relative to their peers, whereas White and Asian/Pacific-Islander females have the strongest performance. Medium size classes have a more varied pattern, but here too, underrepresented males and females experience no change in outcome across all three disciplines. One possible explanation for this pattern might be explained by social group theory. Instructors generally favor smaller class sizes because it allows them to work closely and develop a relationship with their students. However, this reasoning does not consider learning that may happen either between students or even outside of the classroom. Dovidio et al. (2008) and Gonzalez (2000) note that underrepresented students often seek both in-class and out-of-class support from those with whom they most identify and share a similar background, thereby creating a de facto support group that facilitates success in multiple aspects of college life. It stands to reason that in classes with more underrepresented students, there is a higher possibility for such identity groups to be present. However, upon closer inspection of the number of underrepresented students in the two smallest class sizes, these students were enrolled in multiple subject areas with less than 1,000 observations throughout the entire observation period—equating to an average of 14 students per section per term in that discipline.

The findings related to STEM courses also raise some concerns about the long-term implications of the effect on outcomes for student groups who are not well represented in the field—underrepresented minorities and women of all backgrounds. Both appear to be unphased by increased class sizes (save for the decrease experienced by the 2,459 White and Asian/Pacific Islander female students in oversized classes). Conversely, only White and Asian/Pacific Islander males in small STEM classes appear to experience a potential increase in student achievement. Given these opposing estimates, it is essential to remember that the results in Table 5 highlight the correlational effect of class size on a *single* course. Thus, throughout a student's entire GenEd program (and possibly their undergraduate career), there may indeed be a cumulative effect that magnifies inequality in student achievement by race and gender. STEM faculty have historically attributed White and Asian male dominance in these fields to better academic preparation compared to their underrepresented counterparts. However, Griffith (2010) and Riegle-Crumb et al. (2012) have challenged this thinking by arguing that the lack of mentorship and support for underrepresented minorities and women, which would otherwise support retention in undergraduate STEM majors, is sorely lacking in higher education. Thus, as was the case with the previous finding, the lack of a support system for these groups may help explain why student achievement for African American, Hispanic and American Indian students is mostly static regardless of class size when compared to White and Asian/Pacific Islander males.

A third noteworthy finding is that no student group in any of the three disciplines appears to be affected by large size classes. One possible reason for this is course curricular design and instructional delivery. As class size increases, instructors tend to modify the breadth and depth of course objectives, course assignments, and course-related learning outside the classroom (Cuseo, 2007). A class size between 31 and 40 may well be the maximum limit before an instructor is forced to incorporate more time-saving, but less academically meaningful assignments (e.g., curtailing the number of assigned papers or eliminating time-intensive projects) to the detriment of student learning and, ultimately, student achievement. This reasoning is supported by K–12 class size studies in which student academic outcomes tend to either stagnate or decline when classes approach or exceed a "tipping point" generally in the 30s range (Angrist & Lavy, 1999; Boozer & Rouse, 2001), which may also explain why for most students in extra-large and oversized classes (except for the 1,231 White and Asian/Pacific Islander females in the arts & humanities and 2,057 African American, Hispanic, and American Indian females in the social sciences), they appear to experience either no increase in student achievement or an outright decline.

Limitations

There are several notable limitations to the present investigation. First, because this is a correlational analysis, it is not possible to accurately identify the causal effects of class size. Second, although we have provided high school GPA, SAT Math and Verbal scores as controls for prior student ability, our model does not include a control for present student ability such as undergraduate GPA because of the limitation of our database. Third, because R's lme4 function performs a listwise analysis, only 60% of the population is included in the analysis.

Implications and Future Direction

The findings point to some important implications for administrators and faculty as they relate to program evaluation and faculty development. Given the highly variable effects of class size by student race, gender, and academic discipline, it would be challenging to employ a "one-size-fits-all" policy throughout the entire program. Although we can argue that smaller class sizes improve pedagogical and curricular quality, research has shown a more valuable (and perhaps realistic) policy intervention might be to provide instructors with the professional development needed to meet individualized student need (Brancato, 2003; Stes et al. 2010). For instance, because many first-generation minority students often have developing study and time

Table 5
Full Table of Interaction Effects of Model Variables on Transformed GenEd Course Grade ($n = 172.516$)

	Male						Female					
	Overrepresented Groups			Underrepresented Groups			Overrepresented Groups			Underrepresented Groups		
	Social Science GenEds	STEM GenEds	Arts & Humanities GenEds	Social Science GenEds	STEM GenEds	Arts & Humanities GenEds	Social Science GenEds	STEM GenEds	Arts & Humanities GenEds	Social Science GenEds	STEM GenEds	Arts & Humanities GenEds
Small class sizes (up to 25th percentile: ≤20 students)	Reference Group	.114 (.032)***	.003 (.019)	−.148 (.037)***	−.179 (.065)**	−.025 (.038)	.184 (.021)***	−.072 (.037)	.022 (.021)	−.077 (.047)	.056 (.082)	.039 (.048)
n	2,796	961	11,252	515	195	2,053	2,885	1,098	11,303	948	351	3,490
Medium size classes (26th-50th percentile: 26–30 students)	−.020 (.024)	−.025 (.042)	.062 (.027)*	.009 (.046)	.079 (.081)	−.042 (.052)	.062 (.026)*	−.070 (.046)	−.083 (.029)**	.006 (.058)	.001 (.103)	.042 (.065)
n	3,086	1,987	10,725	628	359	1,858	3,516	2,088	10,692	1,205	595	3,249
Large size classes (51st–70th percentile: 31–40 students)	−.025 (.020)	−.048 (.040)	.051 (.039)	.034 (.039)	.039 (.077)	−.123 (.071)	.015 (.022)	−.046 (.043)	.004 (.040)	.003 (.049)	.076 (.098)	.118 (.090)
n	9,376	2,224	1,025	1,804	388	173	9,660	2,304	954	3,160	688	325
Extra large size classes (71st–80th percentile: 41–60 students)	.009 (.023)	−.174 (.047)***	−.077 (.042)	.013 (.043)	.007 (.087)	−.017 (.070)	−.015 (.024)	.029 (.048)	.096 (.039)*	.073 (.054)	−.043 (.109)	−.100 (.087)
n	4,998	1,224	1,122	908	203	204	4,739	1,484	1,231	1,529	400	423
Oversized classes (>80th percentile: >60 students)	−.079 (.025)**	−.033 (.048)	−.004 (.040)	−.053 (.041)	.091 (.077)	−.036 (.054)	.011 (.023)	−.140 (.043)**	−.011 (.030)	.122 (.051)*	−.082 (.098)	−.070 (.070)
n	7,189	3,015	3,207	1,234	446	511	6,700	2,459	2,544	2,057	705	726

Note. Other racial/ethnic groups have been omitted for simplicity.
*$p < .05$. **$p < .01$. ***$p < .001$.

management skills and experience more difficulty navigating institutional bureaucracy, encouraging support groups and expanding access to critical academic resources may help bolster academic performance and retention for these students (Richardson & Skinner, 1992). It is possible then that underrepresented students may need more significant individualized support and attention from their instructors akin to what their overrepresented peers have experienced to date.

Finally, both the findings and sheer amount of data available at the postsecondary level suggest that there are multiple avenues by which the issue of class size can be studied in the future. For example, student evaluation questions, which are provided at the class level in this study, could be used as outcomes in a structural equation model that describes the relationship between course grade, class size, and student perceptions of the course. Additionally, it may be worthwhile to examine which instructional techniques are associated with class size and their overall influence on student learning.

REFERENCES

Angrist, J., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics*, *114*(2), 533–575. https://doi.org/10.1162/003355399556061

Arias, J., & Walker, D. (2004). Additional evidence on the relationship between class size and student performance. *Journal of Economic Education*, *35*(4), 311–329. http://dx.doi.org/10.3200/JECE.35.4.311-329

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review*, *27*(3), 253–265. http://doi.org/10.1016/j.econedurev.2006.08.007

Blatchford, P., Bassett, P., & Brown, P. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: Differences in relation to pupil prior attainment and primary vs. secondary schools. *Learning and Instruction*, *21*(6), 715–730. http://doi.org/10.1016/j.learninstruc.2011.04.001

Boozer, M., & Rouse, C. (2001). Intraschool variation in class size: Patterns and implications. *Journal of Urban Economics*, *50*(1), 163–189. https://doi.org/10.1006/juec.2001.2216

Borland, M., Howsen, R., & Trawick, M. (2005). An investigation of the effect of class size on student academic achievement. *Education Economics*, *13*(1), 73–83. http://dx.doi.org/10.1080/0964529042000325216

Brancato, V. (2003). Professional development in higher education. *New Directions for Adult and Continuing Education*, *2003*(98), 59–66. http://dx.doi.org/10.1002/ace.100

Bryne, B. (2016). *Structural equation modeling with AMOS* (3rd ed.). New York: Routledge

Carbone, E., & Greenberg, J. (1998). Teaching large classes: Unpacking the problem and responding creatively. In M. Kaplan (Ed.), *To improve the academy* (No. 17, pp. 311–326). Stillwater, OK: New Forums Press and the Professional and Organizational Development Network in Higher Education.

Cuseo, J. (2007). The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. *Journal of Faculty Development*, *21*(1), 5–21.

Diette, T., & Raghav, M. (2015). Class size matters: Heterogeneous effects of larger classes on college student learning. *Eastern Economic Journal*, *41*(2), 273–283. http://dx.doi.org/10.1057/eej.2014.31

Dorans, N. (1999). Correspondences between ACT™ and SAT® I scores. *ETS Research Report Series*, *1999*(1). http://dx.doi.org/10.1002/j.2333-8504.1999.tb01800.x

Dovidio, J., Gaertner, S., & Saguy, T. (2008). Another view of "we": Majority and minority group perspectives on a common ingroup identity. *European Review of Social Psychology*, *18*(1), 296–330. https://doi.org/10.1080/10463280701726132

Drewes, T., & Michael, C. (2006). How do students choose a university? An analysis of applications to universities in Ontario, Canada. *Research in Higher Education*, *47*(7), 781–800. http://dx.doi.org/10.1007/s11162-006-9015-6

Ebbes, P., Böckenholt, U., & Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, *58*(2), 161–178. https://doi.org/10.1046/j.0039-0402.2003.00254.x

Figlio, D., Schapiro, M., & Soter, K. (2015). Are tenure track professors better teachers? *Review of Economics and Statistics*, *97*(4), 715–724. http://dx.doi.org/10.1162/REST_a_00529

Gonzalez, K. (2000). Toward a theory of minority student participation in predominantly White colleges and universities. *Journal of College Student Retention: Research, Theory & Practice*, *2*(1), 69–91. https://doi.org/10.2190/LPCF-P0C3-N4BU-464R

Griffith, A. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911–922. https://doi.org/10.1016/j.econedurev.2010.06.010

Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*, 1141–1177.

Hoxby, C. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, *115*(4), 1239–1285. http://doi.org/10.1162/003355300555060

Iran-Nejad, A., McKeachie, W., & Berliner, D. (1990). The multisource nature of learning: An introduction. *Review of Educational Research*, *60*(4), 509–515.

Keil, J., & Partell, P. (1997). *The effect of class size on student performance and retention at Binghamton University*. New York: Binghamton University Office of Budget and Institutional Research.

Kline, R. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guildford Press.

Kokkelenberg, E., Dillon, M., & Christy, S. (2008). The effects of class size on student grades at a public university. *Economics of Education Review*, *27*(2), 221–233. http://doi.org/10.1016/j.econedurev.2006.09.011

Krueger, A. (2003). Economic considerations and class size. *Economic Journal*, *113*(485), F34–F63. http://doi.org/10.1111/1468-0297.00098

Krueger, A., & Whitmore, D. (2002). Would smaller classes help close the black-white achievement gap? In J. Chubb & T. Loveless (Eds.), *Bridging the achievement gap* (pp. 11–46). Washington, DC: Brookings Institution Press.

McKeachie, W. (1980). Class size, large classes, and multiple sections. *Academe*, *66*(1), 24–27. http://doi.org/10.2307/40249328

Monks, J., & Schmidt, R. (2011). The impact of class size on outcomes in higher education. *BE Journal of Economic Analysis and Policy*, *11*(1), 1–17. http://doi.org/10.2202/1935-1682.2803

Podsakoff, P., MacKenzie, S., Lee, J., & Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879.

Richardson, R., & Skinner, E. (1992). Helping first-generation minority students achieve degrees. In L. Zwerling & H. London (Eds.), *First generation college students: Confronting the cultural issues*. San Francisco: Jossey-Bass Publishers.

Riegle-Crumb, C., King, B., Grodsky, E., & Muller, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. *American Educational Research Journal*, *49*(6), 1048–1073. https://doi.org/10.3102/0002831211435229

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Smith, M., & Glass, G. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal*, *17*(4), 419–433.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and applied multilevel analysis*. Thousand Oaks, CA: SAGE.

Stes, A., Coertjens, L., & Van Petegem, P. (2010). Instructional development for teachers in higher education: Impact on teaching approach. *Higher Education*, *60*(2), 187–204. http://dx.doi.org/10.1007/s10734-009-9294-x

Terenzini, P., & Pascarella, E. (1991). Twenty years of research on college students: Lessons for future research. *Research in Higher Education*, *32*(1), 83–92. http://dx.doi.org/10.1007/BF00992835

Wachtel, H. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191–212. http://dx.doi.org/10.1080/0260293980230207

Walia, B. (2008). *Three essays in health and labor economics* (PhD dissertation, Kansas State University, Manhattan, KS).

Williams, D., Cook, P., Quinn, B., & Jensen, R. (1985). University class size: Is smaller better? *Research in Higher Education*, *23*(3), 307–318. http://dx.doi.org/10.1007/BF00973793

Word, E., Johnston, J., & Bain, H. (1990). *Student/Teacher Achievement Ratio (STAR): Tennessee's K-3 class size study. Final summary report 1985-1990*. Nashville: Tennessee State Department of Education.

AUTHORS

ETHAN AKE-LITTLE, PhD, is the executive director of AFT Pennsylvania and Adjunct Instructor at the Temple University College of Education, 1301 Cecil B. Moore Avenue, Philadelphia, PA 19122; *ethanake@temple.edu*. His research focuses on education policy, K–12 teacher retention, and the dynamics of classroom teaching and learning.

NATHANIEL VON DER EMBSE, PhD, is an associate professor of School Psychology, fellow at the Educational Policy Information Center, and codirector of the School Mental Health Collaborative at the University of South Florida College of Education, 4202 E. Fowler Avenue, Tampa, FL 33620; *natev@usf.edu*. His research interests include universal screening for behavioral and mental health, teacher stress and student test anxiety, and training educators in population-based mental health services.

DANA DAWSON, MA, is the associate director of the General Education Program at Temple University, 500 Conwell Hall, Philadelphia, PA 19122; *dgdawson@temple.edu*. Her experience spans over two decades in higher education, including teaching in Temple University's Intellectual Heritage Program as well as launching the university's Office of Scholar Development and Fellowships Advising.

Manuscript received July 13, 2017
Revisions October 9, 2018; November 14, 2019; March 17, 2020
Accepted March 31, 2020