Class Size for a Multilingual Mainstream: Empirical Explorations

Bradley Queen

Abstract

This article presents the findings of an exploratory study that synthesizes local empirical and statistical data to argue about the performances of students, categorized into distinctive linguistic cohorts, in mainstream composition sections of different sizes. The data suggests that section sizes of between fifteen and nineteen enhance the learning environments of all students within a majority multilingual student demographic when portfolio pedagogies are used, a conclusion that offers support ultimately to the policy recommendations on class sizes advocated by the NCTE/CCCC.

Introduction

The NCTE/CCCC position statement on class sizes, "Principles for the Postsecondary Teaching of Writing," recommends that

> Institutions can provide reasonable and equitable working conditions by establishing teaching loads and class sizes that are consistent with disciplinary norms. No more than 20 students should be permitted in any writing class, Ideally, classes should be limited to 15. Remedial or developmental sections should be limited to a maximum of 15 students.

For writing classes with specialized student populations, the "Statement on Second Language Writing and Writers" offers an accompanying set of policies: a cap of twenty students per section "in mainstream classes with a substantial number of second language writers," and a maximum cap of fifteen for classes "made up exclusively of second language writers." All such class sizes can enable "sound writing instruction and reasonable and equitable working conditions" to coexist.

Empirical research on class sizes in postsecondary writing contexts has been hard to come by, as noted by Horning (2007), who could not find "a solid empirical study to demonstrate, once and for all, that smaller classes help students become more effective writers in college" (11). This can present problems for WPAs when using the white paper recommendations in reports to upper administration about learning outcomes and course caps, workloads, and fair labor practices. Without field-specific and local empirical research that critically analyzes the relationship between class sizes and teaching and learning, this dearth of research may contribute to and reaffirm political economies that marginalize composition programs and writing teachers.

This localized study examines class sizes and linguistic demographics in a mainstream gateway writing course that basic writers must pass before registering for general education composition courses.[1] The project gathers a thick body of local evidence—two assessments of portfolios, longitudinal course evaluation data, and an array of descriptive statistics, inferential statistics, and linguistic demographic data—to forward arguments about the performances of students, categorized into distinctive linguistic cohorts, in composition sections of different sizes.

Together these data suggest that in the context of the mainstream writing course studied, section sizes of between fifteen and nineteen enhance the learning environments of all students within a majority multilingual student demographic when portfolio pedagogies are used. Moreover, on the instructional side of the teaching and learning dialectic, such class sizes enable instructional techniques that attend capably to linguistic diversity, benefitting a polyglot student population with different levels of proficiency in written English and distinct reservoirs of cultural and institutional knowledge.

The Local Context: A Multilingual Mainstream

At the University of California, Irvine (UCI), a mainstream composition class looks more like a specialized one where the vast majority of students are multilingual. Writing A, Introduction to Writing and Rhetoric, is the most critically situated of the three writing sequence courses because of the pedagogical exigencies presented by the changing linguistic demographics of its student population. Sitting between the Academic English/ESL Program and two courses that fulfill the lower-division writing requirement, WrA is a vital pedagogical bridge for basic writers.

In recent years, WrA has seen the proportion of its international students expand steadily. In the 2013–2014 academic year, Composition

received permission and funding from the Humanities Dean to experiment with section sizes by lowering the caps on fifteen WrA classes to fifteen students, down from the course cap of twenty. The following fall, the course caps for all three of Composition's courses—A, B, and C—were lowered to nineteen under the auspices of a pilot. The number of sections delivered expanded apace, from 388 in 2013–2014 to 470 in 2014–2015, as did WrA's offerings: 65 sections in 2013–2014, 86 in 2014–2015, and 77 in 2015–2016. By the spring term of 2015, the international student population in WrA reached almost 70%.

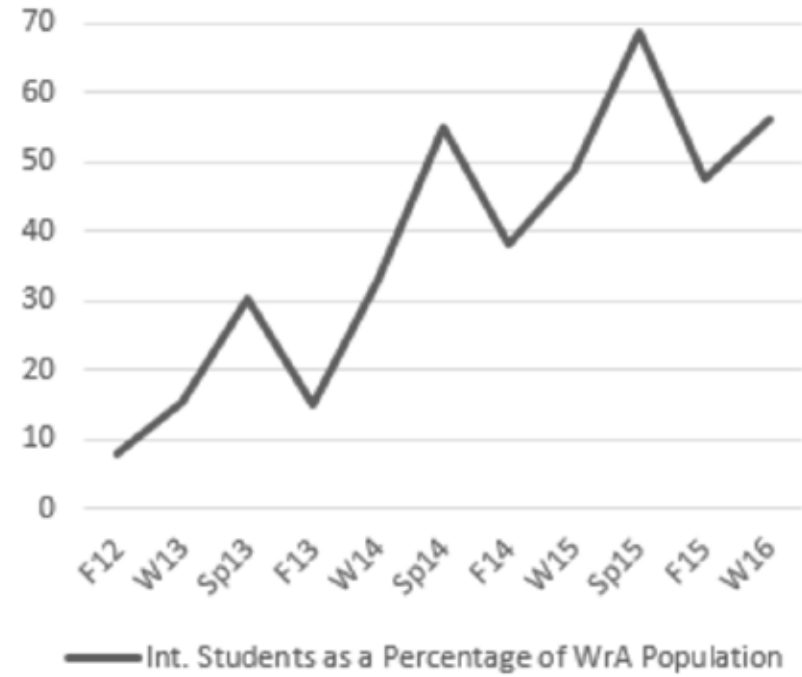
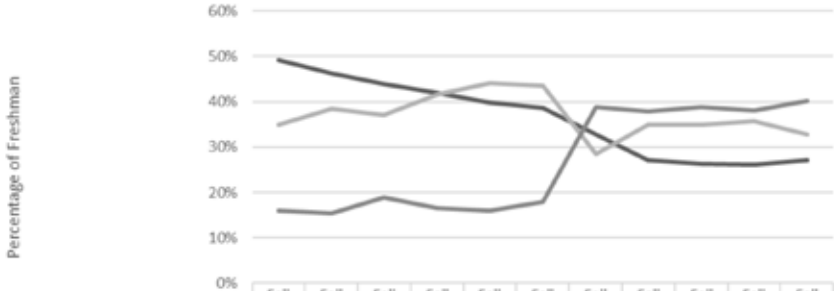


Fig. 1. International Students Population in WrA

Admissions data from across the University of California system documents a significant expansion in the undergraduate international student population. The pattern registers a 28.5% increase from the fall of 2012 to the fall of 2013, another 18% bump in 2014, and another 12.8% in 2015 (UC).

Predictably then the proportion of new undergraduates who report a "primary home language" as "another language only" has also grown markedly over the past few years. When potential UC students fill out the sys-

tem's common application, they come across the About You section. Among various other demographic questions, they find this one, "What language did you learn to speak first?" When the data is processed by admissions offices, responses are given one of three First Language Codes (FLC): 1) English Only, 2) English and Another, or 3) Another Language Only.



	Fall 2005	Fall 2006	Fall 2007	Fall 2008	Fall 2009	Fall 2010	Fall 2011	Fall 2012	Fall 2013	Fall 2014	Fall 2015
FLC 1: English Only	49.1	46.2	43.9	42	39.8	38.6	32.8	27.1	26.3	26.2	27.1
FLC 2: English and Another Language	34.9	38.4	37.1	41.6	44.2	43.5	28.4	35	34.9	35.8	32.8
FLC 3: Another Only	16	15.4	18.9	16.5	16	18	38.8	37.8	38.8	38	40.2

Fig. 2. Entering Freshmen and Linguistic Demographics by Fall Quarter, 2005–2015

At UCI, the proportion of FLC 3 students within the incoming freshmen cohort increased to just over 40% in 2015, up from 16% in 2005, while the English-only proportion has declined, from 49% in 2005 to 27% in 2015. Meanwhile, the proportion of new students who report as FLC 2 has remained steady from 2005 to 2015 at 37% on average (OIR). These shifts affirm trends in US higher education (Ruiz; IIE; Hussar and Bailey).

Literature Review: Factors Impacting the Conversation about Class Size

Factor One: Impact of Class Size on Learning

Over the past decades, researchers throughout academia have argued about class sizes in studies ranging across disciplines and educational levels. The resulting research has been characterized as "broad, diverse, diffuse, and generally unwieldy" (Fleming et al. 7); the findings described as "conflicting, inconclusive, and disappointingly meager" (Robinson and Wittebols 1). Horning's 2007 inability to find the definitive article on class size is echoed in the historical literature of other disciplines. Searching through previous research, particularly in other fields, though, can add to and complicate our understanding of the issue of class size in writing studies. As

far back as 1979, for example, social scientists Glass and Smith noted that "Review after review of the topic has dissolved into cynical despair or epistemological confusion" (2). The fickle nature of the conclusions offered by many studies, so claim compositionists Farrell and Jensen in 2000, still allows "interested parties to support any argument" (322).

Nevertheless, in their first of two meta-analyses, Glass and Smith claim they found "a clear and strong relationship between class size and achievement" (15). In their second, which developed 371 comparisons from 59 studies, they claim "more is learned in smaller classes," with the relationship between class size and achievement slightly stronger at the secondary level than at the elementary (15). However, in 1989 another meta-study by educational economist Eric Hanushek drew from 189 studies and found "no strong evidence that teacher-student ratios . . . have the expected positive effects on student achievement" (47). While the research in elementary and secondary contexts relies on the fundamental hypothesis that smaller class sizes for all students but especially for at-risk populations, language learners, and special needs students can enhance learning and achievement across subjects, the multidisciplinary research on class sizes in postsecondary contexts has focused on questioning the following assumption: "Small class sizes are neither necessary nor sufficient to ensure high quality student learning, growth and development. What matters is not the size of the class but what goes on in the class" (Gilbert 5).

Factor Two: Impact of Institutional/Economic Factors on Class Size

Recent research on class sizes in postsecondary contexts recognizes the difficulties with determining empirically what goes on in classrooms, what factors influence learning environments, and what forms of evidence should be used to analyze teaching and learning. In addition, though, and more explicitly, economic studies make key contributions by theorizing class size as situated within the material financial conditions of local institutions. According to institutional research analyst Steve Chatman, class sizes are tied to "reductionist efficiency goals that consider only production costs" (Chatman 615). Postsecondary institutions attempt to create more efficient ways of delivering instruction and often privilege economic costs and benefits over educational outcomes and the quality of the teaching itself (Cuseo). From this perspective, class sizes become functions of "both institutional and course characteristics with institutional characteristics dominating" (Chatman 628; Correa).

According to economics professors James Monks and Robert Schmidt, such claims can be extended. They argue that class size may be endogenous

to research designs that study its effects on learning and teaching. Existing studies "combine and confound class size effects and total student responsibility effects" (2) where the latter refers to an instructor's overall workload during a given term. Monks and Schmidt use student evaluations to argue that the interpretive correlations between class sizes and learning outcomes are already situated within total workload effects for each instructor and learning environments whose dynamics are influenced by class size as situated.

Such claims have much to tell us about researching class sizes through course-fixed effects, or what might also be called the environmental dynamics of teaching and learning. Regardless of their concerns, ultimately, Monks and Schmidt found that "class size has a negative and statistically significant (at the 95% level or greater) impact on the amount of critical and analytical thinking required in the course, the effectiveness of the teaching methods," and other pedagogical dynamics (13). Class size then does seem to play a role in student achievement, the quality of instruction, and the environmental dynamics of learning itself.

Factor Three: Student Evaluations, Teaching Effectiveness, and Class Sizes

Definitions of effective teaching vary across institutions, fields, disciplines, and class sizes. The association therefore—among the size of a class, effective teaching, and registers of learning—must be situated, perhaps within a particular course in a first-year composition sequence. While student evaluations of teaching seem to offer consistent mechanisms for measurement, standardized processes for determining effective writing instruction do not exist at the postsecondary level. Moreover, many within composition studies—most recently, Adams Wooten, Ray, and Babb in Fall 2016 *WPA: Writing Program Administration*—have taken issue with teaching evaluations as the main body of evidence in a teacher's performance record. While they provide quantitative and qualitative data that can be efficiently compiled and analyzed, evaluations can conflate the formative with the summative, the discussion of writing pedagogies with accountability defined by abstract criteria and opaque procedures. Such criticism has persisted for decades.

Alongside the politics of student evaluations comes research from various academic disciplines documenting the ways student teaching evaluations offer distorted interpretations of a teacher, of teaching effectiveness, or of learning itself. Student evaluations can privilege affective responses over more substantive reflections: how much students like a teacher or whether they were entertained in a particular class can obfuscate more important

determinations of learning and teaching (Centra; Boring et al.). Written comments on evaluation forms can be cruel and riddled with projective anger (Lindahl and Unger). Studies analyzing the influence of gender and gender roles on evaluation results offer mixed and inconclusive findings (Feldman; Freeman; Young et al.). The relationship between race and ethnicity and student ratings needs more analysis within specific disciplines and across institutions (Centra). Inquiries that ask whether student evaluations are influenced by low workloads, leniency in grading, perceptions of expected grades, and levels of rigor are answered only by studies focusing on specific pedagogical contexts (Clayson; Marsh and Roche).

While race, gender, and ethnicity are biasing factors in teaching evaluations, another that needs to be considered is class sizes. There does seem to be a correlation between class size and student evaluations of teaching and learning in postsecondary contexts. As far back as 1993, education professor John Centra argued that studies found biasing relationships in which smaller classes received higher student ratings, especially in "the dimensions of rapport and interaction with students" (qtd. in Feldman 66). While sociologist Kenneth Feldman's meta-analysis did argue that "overall or global ratings of teachers are as likely to be inversely associated with class size as not to be related at all" (66), Feldman took note of pedagogical aspects involving interpersonal interactions to conclude that "two-thirds to three-fourths of the associations (in available studies) between students' ratings on these dimensions and class size are inverse and statistically significant" (Feldman 71–72). The larger the class, generally speaking, the lower the ratings for such aspects of teaching and learning; the smaller the class, the higher the ratings.

Ultimately, Feldman argued that the degree of potential bias as it could be attributed to class size could not be separated from "what it is student ratings measure . . . and how validly they do so" (80). In other words, class size is a fundamental pedagogical element that can correlate with self-reported perceptions of teaching and learning and with learning outcomes themselves (Bedard and Kuhn; Chapman et al.).

When combining Feldman's conclusions with Monks and Schmidt's arguments, a tenable theoretical position emerges to justify using student evaluations as they are here: as key indicators of pedagogical environments within which class size can figure significantly. Current scholarship, in arguing that course evaluations may be unreliable indicators of pedagogical effectiveness, has refocused the use of evaluations for environmental interpretations of teaching and learning and of the pedagogies delivered. Class size, therefore, has become one fundamental aspect among others that can correlate with the performances of students and teachers, and it is among

the major elements defining local classroom environments, in which pedagogical emphases, class level, disciplinary orientation, and overall instructor workload may also be influential. Class sizes and course evaluation data together have become telling evidence when used for formative research within a writing program, especially when put into conversation with other forms of programmatic evidence, such as writing assessment data. The results section reports on such a conversation.

Research Questions, Data Sources, and Procedures

Research Questions

As part of the lowered class size pilot, the composition program was asked to assess the effects of smaller section sizes. In addition to the shifting linguistic demographics, the adoption of a portfolio curriculum, and the transition to a new learning management system, the lowering of course caps added another variable to be studied during an intense moment of curricular reform.

Using the following sources of data, we asked these questions over the course of three academic school years (2013–2014, 2014–2015, and 2015–2016).

1) Do students categorized into institutionally-defined linguistic cohorts perform differently in writing classes of different sizes?

2) Do the conclusions presented in this study support the NCTE/CCCC policy recommendations for class sizes in first-year composition courses?

The data sources are as follows: 1) writing assessment data (see tables 2, 3, 4, and 5); 2) course evaluations (see fig. 3 and table 6); 3) demographic data (see fig. 1, 2, and table 1), and 4) grade data (see fig. 4 and Appendix A).[2]

Procedures

Statistical Analysis: Descriptive and Inferential

The means in all rubric categories from both writing assessments were subjected to F-tests to look for statistical significance in the three linguistic cohorts in classes of different sizes, in the first assessment ≤15 and ≥18, and in the second ≤19. The class sizes for the first assessment were determined by section sizes at the end of terms. The class-size limit for the second was set by the course cap since there was little variability in section size. Interrater agreement counts and percentages from each assessment were compiled and compared, and inferential statistics were compiled for each of the twelve rubric categories for both writing assessments to derive consistency and consensus estimates (Stemler). Pearson and kappa coefficients were

Table 1
Participant Characteristics: First and Second Assessments

Category	Characteristic	Count Assessment 1 n = 258	%	Count Assessment 2 n = 130	%
Sex	Male	100	38.5%	68	52.3%
	Female	150	58.5%	62	47.7%
	Undeclared	8	3.1%		
School, Department, or Program of Primary Major or Affiliation	Dept. of Pharmaceutical Sciences	9	3.5%	2	1.5%
	Information and Computer Sciences	22	8.5%	11	8.5%
	Program in Public Health	9	3.5%	3	2.3%
	School of Biological Sciences	28	10.5%	15	11.5%
	School of Business	6	2.3%	7	5.4%
	School of Education	–	–	1	0.8%
	School of Engineering	25	9.6%	20	15.4%
	School of Humanities	–	–	4	3.1%
	School of Physical Sciences	24	9.2%	10	7.7%
	School of Social Ecology	23	8.9%	4	3.1%
	School of Social Sciences	67	25.8%	26	20.0%
	School of the Arts	6	2.3%	2	1.5%
	Unaffiliated & Undeclared	39	15.11%	25	19.2%

Table 1 (con'd)
Participant Characteristics: First and Second Assessments

Category	Characteristic	Count Assessment 1 *n* = 258	%	Count Assessment 2 *n* = 130	%
Visa	International Student	69	26.7	74	56.9
	Permanent Resident (Includes Undocumented)	28	10.9	18	13.8
	U.S. Citizen	161	62.4	38	29.2
First Language Code (FLC)	FLC 1: English Only	44	17.1	12	9.2
	FLC 2: English & Another Language	71	27.5	29	22.3
	FLC 3: Another Language Only	142	55.0	89	68.5
	Undeclared	1	0.4	0	0
Another Language Only (FLC 3) & Visa	International Student	63	44.4	65	73.0
	Permanent Resident	19	13.4	14	15.7
	U.S. Citizen	60	42.3	10	11.2
English & Another Language (FLC 2) & Visa	International Student	5	7.0	9	31.0
	Permanent Resident	7	9.9	3	10.3
	U.S. Citizen	59	83.1	17	58.6
English Only (FLC 1) & Visa	International Student	0	0.0	0	0
	Permanent Resident	2	4.5	1	8.3
	U.S. Citizen	42	95.5	11	91.7

WPA: Writing Program Administration, Volume 40, Number 2, Spring 2017

Table 2
Comparison of First & Second Assessments: Interrater Agreement, Counts & Percentages

Assessment 1 All Students (*n* = 258)					
	Exact agreement	Adjacent	Scores differ by 2	Scores differ by 3	Scores differ by 4 or more
Number (Percentage)					
Portfolio: Holistic Score	78 (30.2)	113 (43.8)	52 (20.2)	11 (4.3)	1 (0.4)
Portfolio: Rhetorical Awareness	63 (24.4)	128 (49.6)	49 (19.5)	10 (3.9)	0 (0.0)
Portfolio: Self-reflection, Depth & Complexity	65 (25.2)	105 (40.7)	52 (20.2)	9 (3.5)	2 (0.8)
Portfolio: Awareness of Style, Voice & Revision Strategies	65 (25.2)	119 (46.1)	51 (19.8)	14 (5.4)	0 (0.0)
Portfolio: Complexity of Arguments, Persuasiveness & Evidence	72 (27.9)	93 (36.0)	70 (27.1)	14 (5.4)	3 (1.2)
Portfolio: Clarity of Thinking & Expression	69 (26.7)	108 (41.9)	57 (22.1)	16 (6.2)	2 (0.8)
Portfolio: Mechanics, Usage & Awareness of Conventions	83 (32.2)	115 (44.6)	39 (15.1)	12 (4.7)	1 (0.4)
Final Essay: Exposition & Argumentation	75 (29.1)	108 (41.9)	59 (22.9)	15 (5.8)	1 (0.4)
Final Essay: Command of Language & Rhetoric	89 (34.5)	107 (41.5)	48 (18.6)	13 (5.0)	1 (0.4)
Final Essay: Clarity & Readability	86 (33.3)	118 (45.7)	38 (14.7)	11 (4.3)	5 (1.9)
Final Essay: Narrative Development & Control	75 (29.1)	108 (41.9)	52 (20.2)	19 (7.4)	4 (1.6)
Final Essay: Mechanics & Conventions	84 (32.6)	117 (45.3)	44 (17.1)	10 (3.9)	2 (0.8)

Assessment 2 All Students (*n* =130)									
Exact agreement		Adjacent		Scores differ by 2		Scores differ by 3		Scores differ by 4 or more	
	Adj.		Adj.		Adj.		Adj.		Adj.
52 (40)	53 (40.8)	46 (35.4)	75 (57.7)	26 (20.0)	1 (0.8)	6 (4.6)	1 (0.8)	0 (0.0)	0 (0.0)
49 (37.7)	52 (40.0)	56 (43.1)	78 (60.0)	21 (16.2)	0 (0.0)	4 (3.1)	0 (0.0)	0 (0.0)	0 (0.0)
39 (30)	41 (32)	59 (45.4)	83 (64.8)	24 (18.5)	3 (2.3)	7 (5.4)	1 (0.8)	0 (0.0)	0 (0.0)
41 (31.5)	44 (33.8)	60 (46.2)	82 (63.1)	23 (17.7)	2 (1.5)	6 (4.6)	1 (0.8)	0 (0.0)	0 (0.0)
43 (33.1)	46 (35.4)	48 (36.9)	77 (59.2)	31 (23.8)	6 (4.6)	7 (5.4)	1 (0.8)	1 (0.8)	0 (0.0)
43 (33.1)	47 (36.2)	60 (46.2)	77 (59.2)	18 (13.8)	3 (2.3)	8 (6.2)	3 (2.3)	1 (0.8)	0 (0.0)
49 (37.7)	52 (40.3)	62 (47.7)	72 (55.4)	14 (10.8)	5 (3.8)	4 (3.1)	1 (0.8)	1 (0.8)	0 (0.0)
35 (26.9)	35 (26.9)	51 (39.2)	88 (67.7)	33 (25.4)	6 (4.6)	11 (8.5)	1 (0.8)	0 (0.0)	0 (0.0)
55 (42.3)	57 (43.8)	48 (36.9)	69 (53.1)	21 (16.2)	2 (1.5)	5 (3.8)	2 (1.5)	1 (0.8)	0 (0.0)
45 (34.6)	50 (38.5)	59 (45.4)	76 (58.5)	20 (15.4)	2 (1.5)	4 (3.1)	2 (1.5)	2 (1.5)	0 (0.0)
45 (34.8)	46 (35.4)	52 (40.0)	78 (60.0)	21 (16.2)	5 (3.8)	10 (7.7)	1 (0.8)	2 (1.5)	0 (0.0)
57 (43.8)	61 (46.9)	52 (40%)	64 (49.2)	16 (12.3)	4 (3.1)	4 (3.1)	1 (0.8)	1 (.8%)	0 (0.0)

used to compare interrater reliability across the two assessments. Additionally, means from the course evaluation data were subjected to t-tests, and Pearson's product-moment correlation was used to look for statistical significance correlated with changes in section sizes and the pedagogical change to portfolio methods.

Discussion and Data Analysis

Writing Assessment Overview

At the end of the 2013–2014 academic year, final portfolios were selected at random for the first assessment from among the entire body of students who took WrA. A random number table was used. The fifteen sections capped at fifteen were randomly selected and merged with the samples from fifty sections. Each sample was read twice by two of sixteen readers, each an experienced teacher in WrA.

The results were divided by linguistic demographics as defined by the three cohorts of First Language Codes (FLC) within sections of different sizes. When cross referenced with visa codes and with the results of writing assessments, the FLC codes enabled interpretations of each institutionally defined linguistic category (see table 1) across the rubric (see tables 4 and 5).

Comparisons among the FLC cohorts for the first assessment were made within the two class sizes, ≥18 and ≤15. Readers read portfolios remotely. Each of the sixteen readers was given an equal sample set, accessed the rubric online, and filled it out for each sample assigned. No calibration was conducted in the first year other than the weekly staff meetings among the teachers of WrA. No adjudication was performed on the results. Since portfolio pedagogy was new to WrA at the time and its teachers were in the process of adjusting to it and to the shifting linguistic demographics, this first assessment sought to take a preliminary look at WrA's new portfolio curriculum and its attendant methods of judging student work.

A second writing assessment was conducted after the 2016 winter term with samples selected using a random number table. The same rubric from 2013–2014 was used. The five raters for this assessment, again expert WrA teachers, went through a three-hour calibration session prior to conducting the rest of the assessment remotely. An adjudication method was used, and both non-adjudicated and adjudicated scores report agreement and reliability indicators (Collins et al.; Kelly-Riley and Elliot). Adjudication by a third rater, another WrA expert reader, was conducted for scores that differed by two or more on the six-point scale. The third reader read the sample in question, looked at both scores, leveled a third score, and then eliminated a discrepant score (Johnson et al.; Scharf et al.).

Rubric

The rubric categories (see tables 2, 3, 4, and 5) derive from two sources: 1) a faculty survey about pedagogical emphases and 2) common course materials. The rubric comprises two main domains, with the first portfolio and the second final essay. The rubric uses a scale from 5 to 0: Excellent (5), Very Good (4), Good (3), Average (2), Below Average (1), Failing (0). In the portfolio domain, six trait categories follow the holistic evaluation. The final essay domain comprises five traits that evaluate this traditional student product. These two domains embody productive tensions within WrA's evolving curriculum: between portfolio grading and traditional essay grading, and between products and processes as the objects of pedagogical focus, attunement, and assessment.

Interrater Agreement and Reliability

Two statistical forms of interrater agreement are reported for each assessment: descriptive statistics counting the agreement numbers and percentages for each rubric item (see table 2) and inferential statistics reporting consistency and consensus estimates that indicate patterns of agreement (see table 3). Regarding the latter, the kappa statistic measures exact agreement among raters; the weighted form is used to emphasize larger disagreements. The Pearson correlation indicates the consistent nature of patterns of agreements. Together, these empirical registers show reliability as a variegated and multilayered metric for understanding portfolio assessment as a formative heuristic that can reveal productive tensions within a common curriculum.

For WrA's teachers, their recently adopted holistic grading method—in which each teacher provided feedback to their own students for three cycles of process work and drafted essays, and leveled the one course grade via the final portfolio—became a focal point for curricular attunement once the data from the first assessment was disseminated. The left sides of tables 2 and 3 record the results. In terms of the exact agreement percentages, none of the measures are robust. When combined with the adjacent percentages, a number of categories in each domain move above 70%: in the portfolio domain, four of the seven portfolio categories, holistic (74%); rhetorical awareness (74%); awareness of style, voice, and revision (71.3%); and mechanics, usage, and awareness of conventions (76.8%); and in the final essay domain, all five areas, exposition and argumentation (71%), command (76%), clarity and readability (79%), narrative development and control (71%), and mechanics and conventions (77.9%).

WPA: Writing Program Administration, Volume 40, Number 2, Spring 2017

Table 3
Interrater Reliability Indicators

	Assessment 1 *n* = 258		Assessment 2 *n* = 130			
	Non-adjudicated Pearson	Non-adjudicated Weighted Kappa	Non-adjudicated Pearson	Adjudicated Pearson	Non-adjudicated Weighted Kappa	Adjudicated Weighted Kappa
Portfolio: Holistic Score	.45***	.28***	.52***	.76***	.35***	.51***
Portfolio: Rhetorical Awareness	.37***	.19***	.46***	.71***	.30***	.47***
Portfolio: Self-reflection, Depth & Complexity	.42***	.25***	.47***	.72***	.29***	.44**
Portfolio: Awareness of Style, Voice & Revision Strategies	.35***	.19***	.34***	.60***	.17***	.34***
Portfolio: Complexity of Arguments, Persuasiveness & Evidence	.37***	.22***	.45***	.73***	.28***	.46***

Table 3 (con'd)
Interrater Reliability Indicators

	Assessment 1 *n* = 258		Assessment 2 *n* = 130			
	Non-adjudicated Pearson	Non-adjudicated Weighted Kappa	Non-adjudicated Pearson	Adjudicated Pearson	Non-adjudicated Weighted Kappa	Adjudicated Weighted Kappa
Portfolio: Clarity of Thinking & Expression	.32***	.18***	.40***	.66***	.27***	.43**
Portfolio: Mechanics, Usage & Awareness of Conventions	.34***	.22***	.42***	.61***	.29***	.39**
Final Essay: Exposition & Argumentation	.36***	.23***	.39***	.71***	.21***	.41***
Final Essay: Command of Language & Rhetoric	.35***	.23***	.53***	.66***	.31***	.44***
Final Essay: Clarity & Readability	.27***	.20***	.41***	.65***	.24***	.42***
Final Essay: Narrative Development & Control	.31***	.20***	.36***	.70***	.26***	.44***
Final Essay: Mechanics & Conventions	.29***	.18***	.45***	.76***	.34***	.47***

*** $p < .0001$
**$p<.001$

WPA: Writing Program Administration, Volume 40, Number 2, Spring 2017

Table 4 Assessment 1-FLC and Class Sizes

Class Size ≤ 15	N	Mean (5 = Excellent, 2=Average, 0=Failing)	SD	df	F	Class Size ≥ 18	N	Mean	SD	df	F
Portfolio: Holistic											
FLC 1	28	3.5	1.0	2, 141	0.6 (nss)		16	3.6	1.0	2, 107	0.1 (nss)
FLC 2	41	3.6	1.1				30	3.5	1.2		
FLC 3	75	3.4	1.1				64	3.5	1.0		
Portfolio: Rhetorical Awareness											
FLC 1	28	2.7	0.9	2, 136	0.6 (nss)		15	2.8	0.7	2, 107	0.3 (nss)
FLC 2	39	2.7	1.0				30	2.6	1.2		
FLC 3	72	2.5	1.0				65	2.7	0.8		
Portfolio: Self-reflection, Depth & Complexity											
FLC 1	27	2.7	1.0	2, 124	0.3 (nss)		14	2.8	0.9	2, 102	0.2 (nss)
FLC 2	34	2.9	1.0				27	2.9	1.2		
FLC 3	66	2.7	1.0				64	2.7	1.0		
Portfolio: Awareness of Style, Voice & Revision Strategies											
FLC 1	28	2.6	0.9	2, 138	1.2 (nss)		15	2.8	1.1	2, 104	0.8 (nss)
FLC 2	39	2.6	0.9				28	2.4	1.1		
FLC 3	74	2.4	1.0				64	2.4	0.9		
Portfolio: Complexity of Arguments, Persuasiveness & Evidence											
FLC 1	28	2.4	1.2	2, 138	0.4 (nss)		15	2.5	1.3	2, 107	0.1 (nss)
FLC 2	39	2.6	1.1				30	2.5	1.2		
FLC 3	74	2.4	1.1				65	2.4	0.9		
Portfolio: Clarity of Thinking & Expression											
FLC 1	28	2.5	1.0	2, 138	0.9 (nss)		15	2.9	1.1	2, 107	1.1 (nss)
FLC 2	39	2.7	1.0				30	2.5	1.1		
FLC 3	74	2.4	1.0				65	2.5	0.8		
Portfolio: Mechanics, Usage & Awareness of Conventions											
FLC 1	28	2.6	0.9	2, 138	3.0*		15	2.6	0.8	2, 105	5.1**
FLC 2	39	2.6	0.9				28	2.6	0.9		
FLC 3	74	2.3	0.9				65	2.1	0.9		
Final Essay: Exposition & Argumentation											
FLC 1	28	2.4	1.0	2, 142	0.3 (nss)		16	2.4	1.1	2, 109	0.0 (nss)
FLC 2	41	2.5	0.9				30	2.5	1.2		
FLC 3	76	2.4	1.1				66	2.5	1.0		
Final Essay: Command of Language & Rhetoric											
FLC 1	28	2.5	0.9	2, 142	2.1 (nss)		16	2.6	0.9	2, 109	0.5 (nss)
FLC 2	41	2.7	0.9				30	2.6	1.0		
FLC 3	76	2.3	0.9				66	2.4	0.8		
Final Essay: Clarity & Readability											
FLC 1	28	2.8	0.8	2, 142	1.9 (nss)		16	2.8	0.9	2, 109	1.6 (nss)
FLC 2	41	2.8	0.8				30	2.7	0.9		
FLC 3	76	2.5	0.9				66	2.5	0.8		
Final Essay: Narrative Development & Control											
FLC 1	28	2.5	1.0	2, 142	0.4 (nss)		16	2.5	1.0	2, 109	0.3 (nss)
FLC 2	41	2.6	0.9				30	2.4	1.2		
FLC 3	76	2.4	1.0				66	2.6	0.9		
Final Essay: Mechanics & Conventions											
FLC 1	27	2.7	0.7	2, 141	4.5**		16	2.9	0.8	2, 109	5.7**
FLC 2	41	2.7	0.7				30	2.6	0.9		
FLC 3	76	2.3	0.9				66	2.2	0.8		

*p <.05 Note: p values not statistically significant at the 0.05 level are designated as *nss*

**p <.01

Table 5 Assessment 2

Class Size ≤ 19	N	Mean (5 = Excellent, 2=Average, 0=Failing)	SD	df	F
Portfolio: Holistic					
FLC 1	12	3.2	0.7	2, 127	0.2 (nss)
FLC 2	29	3.3	1.2		
FLC 3	89	3.1	1.1		
Portfolio: Rhetorical Awareness					
FLC 1	12	2.3	0.8	2, 126	0.3 (nss)
FLC 2	29	3.3	1.0		
FLC 3	89	3.2	0.9		
Portfolio: Self-reflection, Depth & Complexity					
FLC 1	12	2.1	1.1	2, 125	1.5 (nss)
FLC 2	29	2.6	1.2		
FLC 3	89	2.6	1.0		
Portfolio: Awareness of Style, Voice & Revision Strategies					
FLC 1	12	2.0	0.5	2, 127	1.5 (nss)
FLC 2	29	2.4	0.8		
FLC 3	89	2.5	0.9		
Portfolio: Complexity of Arguments, Persuasiveness & Evidence					
FLC 1	12	1.9	1.0	2, 127	0.9 (nss)
FLC 2	29	2.3	1.1		
FLC 3	89	2.3	1.1		
Portfolio: Clarity of Thinking & Expression					
FLC 1	12	2.1	0.8	2, 127	0.7 (nss)
FLC 2	29	2.5	0.9		
FLC 3	89	2.2	1.1		
Portfolio: Mechanics, Usage & Awareness of Conventions					
FLC 1	12	2.5	0.6	2, 127	4.7**
FLC 2	29	2.6	0.8		
FLC 3	89	2.1	0.9		
Final Essay: Exposition & Argumentation					
FLC 1	12	2.0	1.1	2, 127	0.7 (nss)
FLC 2	29	2.2	1.2		
FLC 3	89	2.3	1.1		
Final Essay: Command of Language & Rhetoric					
FLC 1	12	2.1	0.6	2, 127	0.8 (nss)
FLC 2	29	2.4	0.9		
FLC 3	89	3.2	1.1		
Final Essay: Clarity & Readability					
FLC 1	12	2.4	0.6	2, 127	0.8 (nss)
FLC 2	29	2.6	0.9		
FLC 3	89	2.3	1.0		
Final Essay: Narrative Development & Control					
FLC 1	12	2.2	0.9	2, 127	0.2 (nss)
FLC 2	29	2.3	1.2		
FLC 3	89	2.3	1.0		
Final Essay: Mechanics & Conventions					
FLC 1	12	2.6	0.4	2, 127	5.3**
FLC 2	29	2.6	0.9		
FLC 3	89	2.0	0.9		

**p <.01

Note: p values not statistically significant at the 0.05 level are designated as *nss*

While these percentages may seem moderately strong, none of the rubric categories registers a weighted kappa coefficient that indicates more than a medium level of reliability on the scale recommended by White, Elliot, and Peckham for use when assessing portfolios. This scale's coefficient ranges—specific to portfolios, which are intricate constructs whose complexity may be distorted by scales and "standards obtained from elementalist measures," such as those applied to timed writing placement exams (123)—are slightly less stringent than conventional psychometric scales. The non-adjudicated weighted kappa scale defines the high range as .46–.69, the medium range as .23–.45, and the low range as .1–.22. On this scale, which is more forgiving than an unweighted kappa scale, a number of rubric areas document low consensus agreement numbers (see table 3). In the portfolio domain, five of seven fall here; in the final essay domain, three of five. The remaining kappa numbers for each fall within medium, and they are on the low end of this range. Furthermore, all of the non-adjudicated Pearson coefficients register a medium level of consistent judgment in which high is .48–.71, medium is .23–.47, and low is .1–.22.

In sum, readers found only weak levels of consensus about how the students performed across all rubric traits, and they agreed and disagreed in moderately consistent patterns. Such results suggest this community of teachers had not yet developed a clear and common understanding of the pedagogical values of WrA's holistic method of grading students, in which students are given a single grade for the portfolio and the distinct assignments are not graded.

Catalyzed by such data, three strategies for curricular attunement were recommended.

1) Develop a more clearly articulated sequence of assignments.
2) Consider grading these assignments and the final portfolio separately, and develop a common method for the holistic grading of end-of-term portfolios.
3) Develop a robust sample of portfolios, and discuss them regularly as a group.

In response, the WrA teaching staff accepted the assignment sequence recommendation, framing it as three papers: the first an imitation, the second a genre analysis, and the third a traditional academic essay. The portfolio assignment remained essentially the same. The third suggestion also took. The teachers discussed sample portfolios more regularly in cohort meetings held in addition to regular meetings. This gradual process, in which bottom-up pedagogical strategies evolved alongside formative curricular assessment, helped create cohort acceptance of new curricular emphases and enable flexible adaptations to shifting demographics.

The agreement counts and reliability statistics (see table 2 and table 3) for the second assessment suggest that such attunement strategies also moved WrA teachers closer toward locating common sensibilities for adjudicating the work of their students. Exact agreement numbers before adjudication are marginally higher in every area in both domains except the exposition and argumentation area of the final essay domain. Where formative programmatic assessment is the focus and the writing construct—final portfolios—is more complex and richly textured, such measures need to be used to further discussion among teachers about interpretive methods.

Due to this formative orientation, the second assessment's adjudicated counts and percentages need to be viewed with skepticism. The act of adjudication in this case was used to motivate further discussion about methods of interpreting portfolios, not to bolster reliability measures. Nevertheless, even with the adjudicated counts and percentages bracketed, the inferential statistics support the claim that the curricular attunement strategies set in motion by the first assessment enabled WrA teachers to establish some measure of value-laden common ground.

Consensus is more solid, with six of seven non-adjudicated kappa scores in the portfolio domain now registering within the medium range of .23–.45, and four of five scores within the final essay domain following suit. The consistency estimates for ten of twelve categories rate as medium for non-adjudicated Pearson scores, and the other two rate as high, with the holistic portfolio score as one of the them. All non-adjudicated coefficients for both estimates are highly statistically significant at $p < .001$. Together, the agreement and reliability data from both assessments reveal productive tensions even as they document moderately reliable readings across all rubric categories.

Writing Assessment Data

All three language cohorts register above average performances for the holistic portfolio evaluation with no statistical significance resulting from the comparison among FLC means within the two class sizes for the first assessment (see table 4). Notably, the FLC 2 group records higher means in the portfolio domain and generally outscores the other two in class sizes of ≤15. Nevertheless, no statistical significance results from the comparison among FLC means within the two class sizes in any of the portfolio domain categories except in the area termed *mechanics, usage, and awareness of conventions*. The significance is stronger in the larger section size, but present in the smaller too, as might be expected in this area defined by awareness of traditional aspects of proficiency with written academic English, with those

students designated as ESL students scoring the lowest. While mechanics and usage are not usually considered an area of meta-cognition, in the rubric's portfolio domain, readers analyzed how students discussed and analyzed their awareness of such things.

The final essay domain demonstrates similar patterns of statistical significance in both section size groups. One variable records statistical significance, that of mechanics and conventions. No areas of higher order rhetorical or analytical skill show statistically significant differences among the FLC cohorts within each section size grouping, but all three cohorts perform solidly, with the FLC 2 cohort generally strongest in the smaller size but frequently even with FLC 1 in both sizes. The FLC 1 is generally highest in the larger size and the FLC 3 generally the lowest, but even at moments with each of the other two and highest in the larger size in the final essay's *narrative development and control* category. The only area of statistical significance in the final essay domain for each section size locates within *mechanics and conventions*.

Notably, the FLC 3 cohort attains higher-than-or-equal-to means in the larger class size when compared with the FLC 3 in the smaller in every rubric category except for the *mechanics, usage, and awareness of conventions* area in the portfolio domain and the *mechanics and conventions area* of the final essay domain. Both of these means come in incrementally higher in the smaller section size. When considering that the FLC 2 population in the smaller size scored higher-than-or equal-to in every rubric category than the FLC 2 cohort in the larger sections, these data suggest that the larger classes may focus more studiously on higher order strategies and skills and may have less time to focus on sentence level crafting. But these results are only suggestive, and tests for statistical significance were not run across FLC cohorts within the two class sizes due in parts to the tepid nature of the kappa and Pearson measures for the first assessment and to the lack of statistical significance across the rubric areas in each class size.

The results from the second assessment look similar to the first (see table 5). All three language cohorts perform solidly, but the FLC 2 and FLC 3 groups generally outperform a small FLC 1 cohort. The FLC 2 and FLC 3 cohorts register stronger means in most trait areas in both domains except for the portfolio holistic and the *mechanics, usage and conventions* category, and the *clarity and readability* category in the final essay domain. Only the *usage, conventions, and mechanics* areas have any statistical significance.

The outcomes of both assessments show students in all groups performing well, according to their teachers even as they deliver an inconclusive response to question one: do student categorized into institutionally-defined linguistic cohorts perform differently in writing classes of different

sizes? When considering the policy recommendations queried in question two, these data suggest that the lines drawn between the specialized and mainstream recommendations may be too soft to carry advocacy positions for smaller class sizes in both domains. Moreover, the writing assessment data, which includes the agreement and reliability evidence, indicates that different kinds of data must be used to offer ecological descriptions of class sizes in writing intensive courses and to theorize field-specific measures of portfolio assessment.

Course Evaluations and Class Sizes

Well before the spike in international student enrollment occurs in the fall of 2011, the Composition Program had been tracking student evaluation numbers. In figure 3, all courses from Fall 2007 to Spring of 2014—2,236 sections—document an average overall instructor rating of 6.1 and an overall course rating of 5.9 on a 1 to 7 scale, with 7 at the high end. On the evaluation form, the overall course rating says, "Overall, this course improved my writing," and the overall instructor rating says, "Overall, the instructor is effective, and I would take another class with her/him." While both of the overall ratings demonstrate solid numbers, the overall instructor rating noticeably averages higher than the overall course rating cumulatively, suggesting students may prefer the personalities of the instructors and their ways of comporting over the learning environment and the substance of the pedagogies.

This relationship looks a bit different when considering the 2014–2015 and 2015–2016 academic years, the first two years in which all of Composition's sections had max caps of nineteen students. Both measures tick up, with the instructor rating moving up by 0.3 and the course rating by 0.4. The relationship looks different yet again when considering the fifteen sections of WrA delivered during 2013–2014. These sections, with caps of fifteen, call into question the pattern between the overall indicators. The course rating comes in at 6.5, incrementally higher than the instructor rating of 6.4. While the census method used may offer only broad generalizations, these small ticks suggest that WrA students in classes capped at fifteen highly valued their curricular environments, within which class size seems to figure significantly.

To make fine-tuned interpretations of the relationship between class sizes and course evaluation numbers, tests were run for statistical significance (see table 6). The first comparison drew from five measures on evaluations gathered for all classes from Fall 2007 to Spring 2015 with enrollment sizes of ≤19 and ≥21. All means for the smaller sections came in higher, and all traits show high statistical significance.

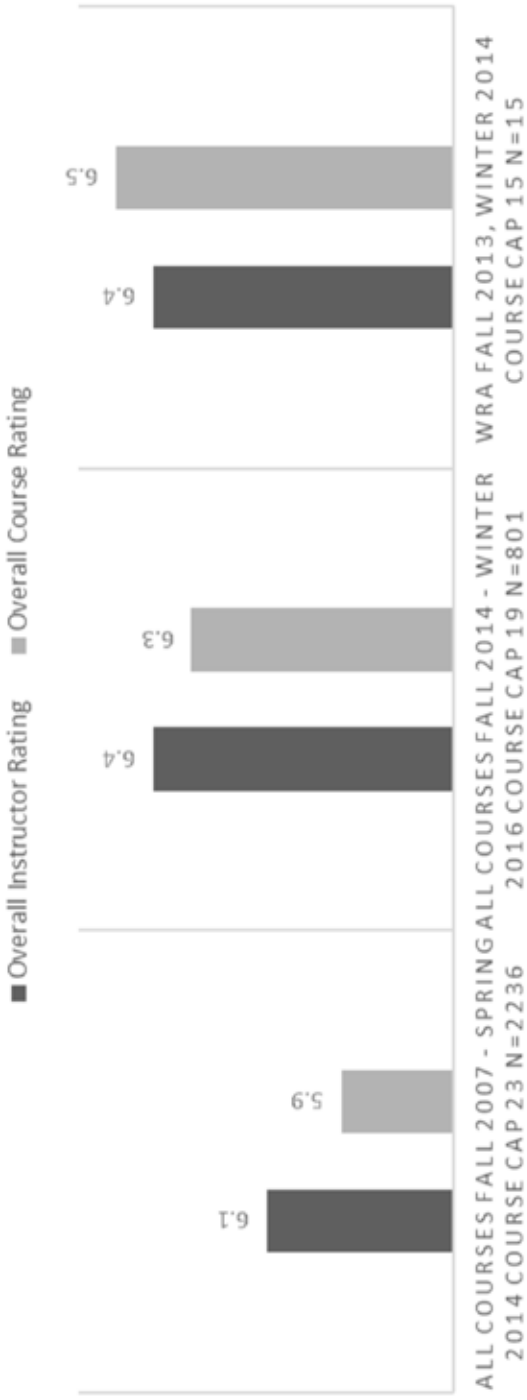


Fig. 3. Overall Course Ratings, Overall Instructor Ratings, and Course Caps

Table 6. Comparison of Evaluation Traits

All Classes	Class Size Comparison							
	≤ 19			≥ 21				
	N	M	SD	N	M	SD	t	df
Enthusiasm	865	6.5	0.4	1700	6.4	0.4	4.7***	2563
Interest	865	6.0	0.6	1700	5.9	0.6	6.9***	2563
Open & Fair	865	6.5	0.4	1700	6.4	0.4	2.9**	2563
Overall Instructor Rating	865	6.2	0.7	1700	6.1	0.7	3.9***	2563
Overall Course Rating	865	6.2	0.6	1700	5.9	0.6	10.9***	2563
	≤ 15			≥ 18				
	N	M	SD	N	M	SD	t	df
Enthusiasm	144	6.6	0.5	2602	6.5	0.5	1.9 (nss)	2744
Interest	144	6.0	0.7	2602	5.9	0.6	2.4*	2744
Open & Fair	144	6.5	0.5	2602	6.5	0.4	0.8 (nss)	2744
Overall Instructor Rating	144	6.2	0.7	2602	6.1	0.7	1.2 (nss)	2744
Overall Course Rating	144	6.1	0.6	2602	5.9	0.6	2.8**	2744

Evaluation Traits WRA								
	≤ 19			≥ 21				
	N	M	SD	N	M	SD	t	df
Enthusiasm	153	6.7	0.3	82	6.3	0.5	6.2***	105.3
Interest	153	6.3	0.4	82	5.7	0.6	8.6***	109.6
Open & Fair	153	6.7	0.2	82	6.4	0.4	6.2***	109.3
Overall Instructor Rating	153	6.5	0.4	82	5.7	0.7	5.8***	109.2
Overall Course Rating	153	6.5	0.3	82	5.9	0.5	8.9***	114.9
	Pre Portfolio			Post Portfolio				
	N	M	SD	N	M	SD	t	df
Enthusiasm	140	6.4	0.5	204	6.7	0.2	-6.6***	186.2
Interest	140	5.9	0.6	204	6.4	0.4	-9.4***	204.8
Open & Fair	140	6.4	0.4	204	6.7	0.2	-6.3***	195.0
Overall Instructor Rating	140	6.1	0.7	204	6.6	0.4	-6.3***	196.1
Overall Course Rating	140	6.0	0.5	204	6.5	0.3	-9.9***	197.9

*$p<.05$

**$p<.01$

***$p<.001$

Note: p values not stastistically significant are designated as *nss*

WPA: Writing Program Administration, Volume 40, Number 2, Spring 2017

The second comparison involves enrollment sizes at ≤15 and ≥18. All means for the classes with the smaller population are higher. Here, just two categories—stimulated interest and overall course rating—show some statistical significance. Together, the comparisons suggest that writing sections with caps lower than nineteen influence positively both the students' affective perceptions of their teachers and their impressions of their learning environments. They also suggest that sizes of between fifteen and nineteen and beneath fifteen may not influence significantly the students' impressions of their teachers, but sections of such sizes may enhance learning environments and influence positively students' perceptions of course-specific pedagogies.

When applied to just WrA, comparisons of sizes at ≤19 and ≥21 yield similar results. All means for the smaller class size are higher and highly statistically significant. The Pearson correlation as applied to all available WrA sections (n = 344) shows that the stimulated interest category (r = -.335) and the overall course rating (r = -.376) register high statistical significance (p <.001) in the pre/post portfolio pedagogy comparison. Together, these data suggest that the larger a section becomes, the less interested the students become and the less they appreciate their learning environments; the smaller it becomes, the more interested the students become in the course and the more they value their learning environments. These conclusions, however interesting they seem, should be viewed as tentative due to the lack of statistical significance demonstrated by the writing assessment data. Nevertheless, as environmental registers, the course evaluation data offers solidly suggestive evidence about learning environments, class sizes, and portfolio pedagogy.

One might argue that grade inflation influences the upward movement of these numbers. The relationship between grades, whether actual or anticipated, and evaluation scores has been problematic for researchers. Some claim higher grades along with higher evaluations reflect greater learning, others that such a situation reflects a lack of rigor (Smith et al.; Krautmann and Saunder). Some conclude "that course evaluations do indeed reflect student learning" when considering specific pedagogical environments (Beleche et al. 710, 718; Clayson, 26–27). This relationship describes the validity hypothesis, which claims that students who have learned more receive higher grades or the grades the feel they deserve and "will naturally rate the professor more highly because of the knowledge they have gained in the course" (Patrick 241). When students think they have learned something, they give course-related and environmental elements ratings that can register as statistically significant, as we have seen thus far.

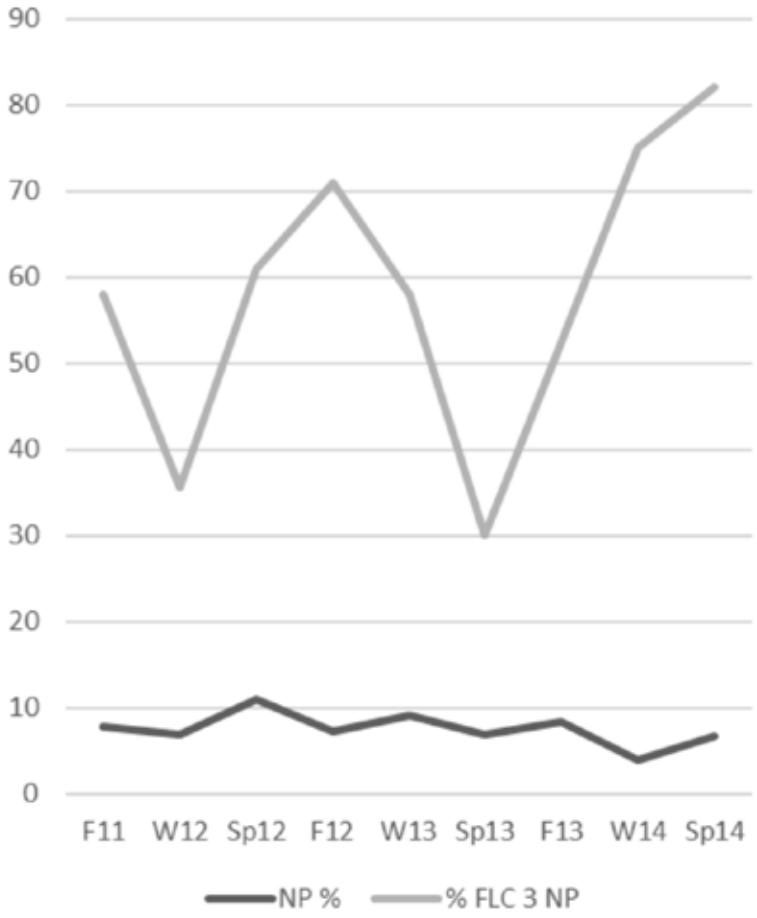


Fig. 4. Non-pass Rate and FLC 3

In WrA, evaluation numbers have gone up even as grades for its FLC 3 majority have trended downward. The proportion of FLC 3 students within the non-pass rate has increased, as the non-pass rate has remained consistent (see fig. 4), and the middle categories of the grade ranges for this cohort show only somewhat discernable patterns (see Appendix A). Over time, this cohort claims increasingly larger proportions of the B/B- and C+/C groups, a downward trend.

These patterns demonstrate neither grade inflation nor grade deflation, either of which might be expected in response to changing demographics, the adoption of portfolio methods, and shrinking class sizes. The grade data suggests that anticipated or actual grades are not biasing factors when WrA

WPA: Writing Program Administration, Volume 40, Number 2, Spring 2017

students rate instructors and learning environments. Smaller class sizes seem to enhance learning environments for WrA's polyglot mainstream.

Conclusion

In conclusion, and to return the study's research questions, the writing assessment data suggests multiple interpretations. One answer is yes: students within distinctive linguistic cohorts perform differently in proficiency-driven categories within classes of different sizes. The FLC 3 students, as expected, perform lower in all the section sizes in the proficiency areas in each of the two domains on the rubric. The FLC 3 cohort in the smallest section sizes performed incrementally better than the FLC 3 cohorts in the larger sizes in the proficiency areas. Another answer is no: the cohorts do not perform differently in classes of fifteen-to-nineteen students in all other categories. Within the rest of the rubric categories patterns in performance results cannot be discerned, and little statistical significance was discovered. Such evidence suggests that smaller class sizes, accompanied by portfolio pedagogies, enable teaching strategies in mainstream writing classes that attend capably to students with widely divergent proficiency levels in written English.

The synthesis presented here—of environmental data that taps learning's affective domains with performance data that emphasizes its cognitive domains—would be untenable without theoretical advances in construct validity. With validity reconceived over the past few decades as argumentation (Huot; Moss), writing studies scholars have moved to problematize psychometric definitions of reliability, with the reliability scale used here as a primary example (White et al.; Elliot et al.). Such developments enable researchers to document the rich complexity of local writing constructs—student portfolios in this study—which provide the evidentiary foundation for the material synthesis where assessment theory integrates with local empiricism to engage pressing field-specific questions.

Ultimately, this exploratory study supports the policies on class sizes advocated by the NCTE/CCCC, even if it does not confirm the distinctions between the recommendations for mainstream and specialized classrooms. Such separate recommendations may not only contradict each other but may also perpetuate discriminatory barriers for multilingual students and stymie pedagogical reforms that should attend to shifting linguistic demographics. In sum, the evidence presented suggests that class sizes of between fifteen and nineteen for basic and first-year writing do foster sound learning environments in which a multilingual student population defines the mainstream writing classroom.

Notes

1. This study has been approved by IRB (HS# 2015-2208).

Appendix A: FLC 3 Grade Distribution

https://goo.gl/zovybD

Works Cited

Adams Wooten, Courtney, Brian Ray, and Jacob Babb. "WPAs Reading SETs: Toward an Ethical and Effective Use of Teaching Evaluations." *WPA: Writing Program Administration*, vol. 40, no 1, 2016, pp. 50–66.

Bedard, Kelly, and Peter Kuhn. "Where Class Size Really Matters: Class Size and Student Ratings of Instructor Effectiveness." *Economics of Education Review*, vol. 27, no. 3, 2008, pp. 253–65.

Beleche, Trinidad, David Fairris, and Mindy Marks. "Do Course Evaluations Truly Reflect Student Learning? Evidence from an Objectively Graded Post-Test." *Economics of Education Review*, vol. 31, no. 5, Oct. 2012, pp. 709–19.

Boring, Anne, Kellie Ottoboni, and Philip B. Stark. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research*, Jan. 2016.

Centra, John A. *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*. Jossey-Bass, 1993.

Chapman, Lauren, and Larry Ludlow. "Can Downsizing College Class Sizes Augment Student Outcomes? An Investigation of the Effects of Class Size on Student Learning." *JGE: The Journal of General Education*, vol. 59, no. 2, 2010, pp. 105–23.

Chatman, Steve. "Lower-Division Class Size at U.S. Postsecondary Institutions." *Research in Higher Education*, vol. 38, no. 5, Oct. 1997, pp. 615–30.

Clayson, Dennis E. "Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature." *Journal of Marketing Education*, vol. 31, no. 1, April 2009, pp. 16–30.

Collins, Regina, Norbert Elliot, Andrew Klobucar, and Fadi P. Deek. "Web-based Portfolio Assessment: Validation of an Open Source Platform." *Journal of Interactive Learning Research*, vol. 24, no. 1, 2013, pp. 5–32.

Conference on College Composition and Communication. "CCCC Statement on Second Language Writing and Writers." Nov. 2014.

—. "Principles for the Postsecondary Teaching of Writing." March 2015.

Correa, Hector. "An Economic Analysis of Class Size and Achievement in Education." *Education Economics*, vol. 1, no. 2, 1993, pp. 129–35.

Cuseo, Joe. "The Empirical Case Against Large Class Size: Adverse Effects on the Teaching, Learning, and Retention of First-Year Students." *Journal of Faculty Development*, vol. 21, no. 1, 2007, pp. 5–21.

Elliot, Norbert, Andre A. Rupp, and David M. Williamson. "Three Interpretive Frameworks: Assessment of English Language Arts-Writing in the Common Core State Standards." *Journal of Writing Assessment*, vol. 8, no. 1, 2015.

Farrell, Edmund J., and Julie M. Jensen. "Rhetoric and Research on Class Size." *Perspectives on Writing: Research, Theory, and Practice*, edited by Roselmina Indrisano and James R. Squire, International Reading Association, 2000, pp. 307–25.

Feldman, Kenneth A. "Class Size and College Students' Evaluations of Teachers and Courses: A Closer Look." *Research in Higher Education*, vol. 21, no. 1, 1984, pp. 45–116.

Fleming, Thomas, Tara Toutant, and Helen Raptis. "Class Size and Effects: A Review." *Phi Delta Kappa Educational Foundation*, 2002. eric.ed.gov/?id=ED478538.

Freeman, Harvey R. "Student Evaluations of College Instructors: Effects of Type of Course Taught, Instructor Gender and Gender Role, and Student Gender." *Journal of Educational Psychology*, vol. 86, no. 4, 1994, pp. 627–30.

Gilbert, Sid. *Quality Education: Does Class Size Matter?* Canadian Society for the Study of Higher Education, 1995.

Glass, Gene, and Mary Lee Smith. *Meta-Analysis of Research on the Relationship of Class Size and Achievement*. Far West Laboratory for Educational Research and Development, 1978.

—. "Meta-Analysis of Research on Class Size and Achievement." *Educational Evaluation and Policy Analysis*, vol. 1, no. 1, Jan./Feb. 1979, pp. 2–16.

Hanushek, Eric A. "The Impact of Differential Expenditures on School Performance." *Educational Researcher*, vol. 18, no. 4, May 1989, pp. 45–62.

Horning, Alice. "The Definitive Article on Class Size." *WPA: Writing Program Administration*, vol. 31, no. 1/2, Fall/Winter 2007, pp. 11–34.

Huot, Brian. *(Re)Articulating Writing Assessment for Teaching and Learning*. Utah State UP, 2002.

Hussar, William, and Bailey, Tabitha. "Projections of Education Statistics to 2022." U.S. Department of Education, U.S. Government Printing Office, 2013.

Institute of International Education (IIE). "International Students: Enrollment Trends." *Institute of International Education*, www.iie.org/Research-and-Publications/Open-Doors/Data/International-Students/Enrollment-Trends.

Johnson, Robert, Jim Penny, Steve Fisher, Therese Kuhs. "Score Resolution: An Investigation of the Reliability and Validity of Resolved Scores." *Applied Measurement in Education*, vol. 16, no. 4, 2003, pp. 299–322.

Kelly-Riley, Diane, and Norbert Elliot. "The *WPA Outcomes Statement*, Validation, and the Pursuit of Localism." *Assessing Writing*, vol. 21, 2014, pp. 89–103.

Lindahl, Mary W., and Michael L. Unger. "Cruelty in Student Teaching Evaluations." *College Teaching*, vol. 58, no. 3, 2010, pp. 71–76.

Krautmann, Anthony, and William Saunder. "Grades and Student Evaluations of Teachers." *Economics of Education Review*, vol. 18, no. 1, 1999, pp. 59–63.

Marsh, Herbert, and Lawrence A. Roche. "Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias, Validity, or Innocent Bystanders?" *Journal of Educational Psychology*, vol. 92, no. 1, 2000, pp. 202–28.

Monks, James, and Robert. Schmidt. "The Impact of Class Size and Number of Students on Outcomes in Higher Education." *Cornell Higher Education Research Institute*, 2010, digitalcommons.ilr.cornell.edu/workingpapers/114/.

Moss, Pamela A. "Comments on Lissitz and Samuelsen Reconstructing Validity." *Educational Researcher*, vol. 36, no. 8, 2007, pp. 470–76.

Office of Institutional Research (OIR). "By Primary Home Language." *University of California, Irvine,* 2015, www.oir.uci.edu/.Accesssed 1 March 2016.

Patrick, Carol Lynn. "Student Evaluations of Teaching: Effects of the Big Five Personality Traits, Grades and the Validity Hypothesis." *Assessment & Evaluation in Higher Education*, vol. 36, no. 2, March 2011, pp. 239–49.

Robinson, Glen E., and James H. Wittebols. *Class Size Research: A Related Cluster Analysis for Decision Making*. Educational Research Service, 1986.

Ruiz, Neil G. "The Geography of Foreign Students in U.S. Higher Education: Origins and Destinations." The Brookings Institution, 2014, https://www.brookings.edu/interactives/the-geography-of-foreign-students-in-u-s-higher-education-origins-and-destinations. Accessed 1 March 2016.

Scharf, Davida, Norbert Elliot, Heather A. Huey, Vladimir Briller, and Kamal Joshi. "Direct Assessment of Literacy Using Writing Portfolios." *The Journal of Academic Librarianship*, vol. 33, no. 4, July 2007, pp. 462–78.

Smith, Dale, Patrick Cook, and William Buskist. "An Experimental Analysis of the Relation between Assigned Grades and Instructor Evaluations." *Teaching of Psychology*, vol. 38, no. 4, October 2011, pp. 225–28.

Stemler, Steven E. "A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability." *Practical Assessment, Research & Evaluation*, vol. 9, no. 44, January 2004, http://pareonline.net/. Accessed 15 March 2015.

"UC Student/Workforce Data." University of California Office of the President, 2015, www.ucop.edu/institutional-research-academic-planning/content-analysis/ug-admissions/student-workforce-data.html.

White, Edward M., Norbert Elliot, and Irvin Peckham. *Very Like a Whale: The Assessment of Writing Programs*. Utah State UP, 2015.

Young, Suzanne, Leslie Rush, and Dale Shaw. "Evaluating Gender Bias in Ratings of University Instructors' Teaching Effectiveness." *International Journal for the Scholarship of Teaching and Learning*, vol. 3, no. 2, July 2009, http://digitalcommons.georgiasouthern.edu/ij-sotl/vol3/iss2/19/. Accessed 2 February 2016.

Acknowledgments

The author would like to thank the editors of *WPA: Writing Program Administration*, the anonymous reviewers, and especially Norbert Elliot for their time and insights. A very special thank you to Anita Iannucci, statistician extraordinaire, and to my colleagues in the English Department, particularly Daniel M. Gross.

Bradley Queen is Associate Director of the Composition Program at the University of California, Irvine, one of ten public, research-intensive universities in the UC system. He also currently serves as chair of the University Committee on Preparatory Education, a systemwide body that evaluates entry gateways and college preparedness policies. He is currently at work on a project studying knowledge transfer strategies of multilingual students across a first-year composition sequence.