Research Article

Divergent Writer Development in College

David Eubanks, *Furman University*

Sara Vanovac, *Furman University*

Structured Abstract

- **Background:** The study uses a novel survey of faculty to collect observational assessments of student writers, which accumulated tens of thousands of samples over five years at a small private liberal arts college. The study was designed to better understand student development over time, motivating the need for large representative sample sizes. The resulting data set of ratings, augmented by student characteristics, was analyzed using hierarchical methods that extend Rasch-type methods by introducing the explanatory variables time (to measure development) and general academic ability.

  The reliability and validity of the data are analyzed, and the results are contextualized with national data sets.

- **Literature Review:** The literature review focuses on validity arguments, observational assessment, and writer theory and assessment. The validity discussion relies on Kane (1992) and Moss (1994) to establish validity as an interpretive argument extending from claims to evidence for and against those claims, to the ethical consequences of use. Kane's (2010) advice on fairness is considered with Poe and Elliot's (2019) applications to writing assessment.

  We find support for observational assessment in K-12 educational research, including a meta-analysis (Südkamp et al., 2012).

  The methods of data gathering and analysis are situated within the tradition of writing assessment using Behizadeh and Engelhard's (2011) historical review

to illustrate how the observational ratings and hierarchical analysis are natural extensions of that tradition.

The finding of an outcome divergence (a "Matthew effect" in Stanovich, 2009) in writing development is contextualized within higher education's other paths of divergence, including standardized test scores and post-graduation income.

- **Research Questions:** Because the observational ratings are easier to obtain than rubric-rating individual papers, the method offers an economical way to assess writing development. The most important research question is to ask if the data have suitable statistical properties to be used as measurements in the aggregate. An affirmative answer would be supported by novel findings concerning student development in writing, such as what student characteristics predict increases over time. In short, does the data collection seem to produce usable data, and if so, what can we learn from it about writer development in college?
- **Methodology:** Data collection is from an end-of-term survey of faculty in a format similar to reporting course grades (one rating per student, per course, per learning outcome). Instructors are asked for an overall assessment of each student as a writer when there is a basis for a judgment. This differs from the more common method of assessing writing samples, for example, using rubrics to rate individual work samples or portfolios. First, the ratings depend on a holistic assessment of a student's work in a term rather than a single paper in isolation. Second, the ratings are fully contextual, rather than being judged by a disinterested (and decontextualized) third party. Because the ratings are intended to show change over time, we call them developmental assessments of student learning, referred to throughout as DASL.

  Additional explanatory information comes from student grades and freshman survey items that provide self-evaluations of writing, mathematics, and other skills and traits. Registration data links ratings to students and their seniority and to the subject and level of coursework associated with the ratings. A relative grade average is calculated as a predictor, omitting classes that are confounded with assessment ratings, and adjusted for course difficulty.

  We use hierarchical (mixed effects) models that are intended to identify overall trends in the data while accounting for rater styles, student individuality, and discipline-specific variance. The regression model is applied twice: once assuming the outcome variable is scalar and once assuming it is ordinal. Model fit is assessed statistically and visually by graphing empirical and modeled outcomes.

For external comparison, we used a national data set of student survey items from the Higher Education Research Institute (HERI) to compare grades to self-assessed writing ability from paired samples during the first and final years of an undergraduate degree.

- **Results:** Comparisons between DASL faculty ratings of students for eight learning outcomes using factor analysis show reasonable clustering by type, including a broad dimension for writing. A similar analysis of student self-ratings likewise shows a factor that associates naturally with writing. Comparisons between self-ratings and instructor ratings show that academic or writing self-confidence predicts writing ratings more than other self-reported traits.

  The reliability of individual faculty ratings is fairly low by testing standards (the intraclass correlation is about .47), but this is easily made up for by the volume of data produced.

  The evidence shows that average DASL ratings plausibly measure a broad construct of writing and show a divergence over time in measured performance levels: students with higher grade averages start higher and gain faster than those with low grade averages. Similar divergence patterns are found independently in national data sets on writer attitudes, standardized tests, and post-graduation incomes.

  The results suggest that outcome divergence is common in higher education, posing questions about the relative opportunities and fairness of a college education.

- **Discussion:** A validity framework structures the analysis of results. Claims are compared and contrasted with the evidence to identify weak points in the argument. Threats to validity include "social promotion" ratings that may inflate scores, and score compression, which could artificially produce diverging averages.

  We claim that the answer to both research questions is yes when the data are aggregated: the averages of the ratings seem to meaningfully associate student characteristics with writing development. Students with lower grade averages have dramatically different developmental curves than students with high grade averages. Because this trend is independently found in related data, the results are likely to be generalizable to other institutions.

- **Conclusions:** The literature on the Matthew effect suggests that skill divergence is caused by feedback mechanisms that produce, maintain, and widen average achievement gaps over time. The results lead us to ask if university structures unwittingly produce and maintain learning structures that

are unfair to subpopulations of students. The regression analysis points to the understanding of grading, both the process of marking and the causes of academic performance, as a key to understanding skill divergence over time.

Keywords: institutional assessment, learning outcomes, rubrics, validity, writing analytics

1.0 Background

White et al. (2015) describe a historical barrier to understanding writing quality: "In cases involving very large samples, it is simply impossible to provide performance information on every student . . . " (p. 115). This sampling problem is due to the difficulty of balancing the mass production of ratings with assuring the quality of the ratings. But what if there were an *easy* way to collect assessments of student writing? What if, instead of working very hard to get 200 samples, we could more easily get 2,000 or 20,000? This study demonstrates that this can be done if we change perspectives: instead of asking about the quality of student *writing*, ask about the qualities of students *as writers*. What we give up and what we gain is the subject of this study. In the end, we will show that the method does work and that the results paint a troubling picture of learning in college.

To gather data, course instructors were asked to subjectively rate students as writers at the end of a term if they had a basis for judgment. These "developmental assessments of student learning" (DASL) ratings employ a holistic rubric designed to detect change over time, and the data collection process is so painless to instructors that 40 percent of them provide data with no more encouragement than an invitation email.

This study analyzes 18,536 DASL ratings of 4,774 unique undergraduate student writers, an average of 3.9 ratings per student. The ratings were collected from Fall 2015 through Fall 2019 each fall and spring semester at Furman University, a selective private liberal arts college with about 2,700 undergraduates and 240 full-time faculty, of whom 98 percent hold terminal degrees in their field. In 2016, the university received the Council for Undergraduate Research's award for excellence. These qualifications show that faculty are suitable for making judgments about student writing in the style of their respective disciplines. Ratings of student writing "in the discipline" were assigned by 289 unique course instructors who had the opportunity to observe the students as writers during a fall or spring course (about fifteen weeks). Most class sections (66%) have fewer than twenty students, and the curriculum is designed to promote out-of-class interactions between students and the faculty.

The intent of this paper is to give grounding, data, and results for a method of assessing writers that is based on observation and natural descriptive language. The primary question to answer is: are we just collecting random numbers? If not, the data collection practice may be of general interest as a complementary approach to other methods.

Because this paper is primarily a validity study, the data and results sections are dense. For readers wanting a quick overview of the findings, we recommend starting with Figure 4, which

graphs average developmental growth in writing, as assessed by the DASL ratings. Students with lower grades start lower and grow slower. Figure 6 shows that the same trend is true among student self-ratings of writing in a large historical data set. Figure 7 shows that it is also true in standardized test scores of critical thinking and writing. Figure 8 shows the divergence in post-graduation incomes. With that background, one can skip to the conclusions in section 7.0 for an overview of findings and recommendations.

2.0 Literature Review

Since the intent of this paper is to demonstrate the usefulness of a novel data-gathering method in writing assessment, we have only included here the parts of the literature review that are needed to understand what is new. Other sources are integrated into the exposition when they are needed, as with the Matthew effect.

The need for new approaches to assessment was nicely summed up by Anderson et al. (2015), who reviewed the impact of writing on learning and described limitations of common methods that "suggest the value of identifying new, more generalizable variables for studying writing's contribution to learning across the entire curriculum" (p. 203). We argue that this project begins with assessing writing itself across the entire curriculum and focuses "specifically on one population with many members from which a large, representative sample can be obtained" (Anderson et al., 2015, p. 203).

2.1 Validity

We rely on Kane's (1992, 2006) validity framework that includes ethics and Moss' (1994) challenge to the primacy of standardized testing by emphasizing hermeneutics. More recently, the distinction between validity and fairness has been the focus of theoretical work generally (Kane, 2010) and within writing assessment (Poe & Elliot, 2019). We take from these works the following guiding principles: (1) validity is an interpretive argument, not a fact to be proved; (2) findings are contextual, not universal; (3) beliefs must be tested for plausibility as well as falsification using comparative and contrasting evidence; and (4) the ethical consequences of belief must be weighed against the evidence. Because fairness depends on the properties of data with respect to vulnerable populations as well as how the findings are used, that discussion is held until the concluding sections.

Validity is taken up in the discussion section, organized by the outline of Kane's (1992, 2006, 2010) validity framework found in White et al. (2015). The framework has the following components:

- scoring (studies of consistency and accuracy),
- generalization (studies of the specific scores in relation to the universe of generalization),
- extrapolation (studies of the relationship between the construct at hand and related constructs), and

- implications (studies of consequences).

See also Cook et al. (2015) for a similar description.

For our purposes, generalization and extrapolation are so intertwined that we treat them together. The last step (implications) includes considerations of fairness which envision validity broadly overlapping with fairness, asking, "Are the proposed interpretations and uses of the test scores appropriate for a population over some range of contexts?" (Kane, 2010, p. 177).

2.2 Observational Assessment

To describe student achievements, we must first observe them. Commonly, the subject of study is a paper a student has written. This is satisfyingly empirical since a paper is a fixed object of study: it looks like a scientific investigation. By contrast, the DASL ratings are assessments of writing samples over a semester of work that may also capture writing behaviors that an instructor will be aware of, such as the difficulty a student had with an assignment. There is a distinction between perceived ability and actual performance in a particular instance; we are asking for the former.

Such behavioral ratings seem to be more common in elementary school settings than in higher education assessment. For example, the observation of children learning to read (Casbergue, 2010) leads to insights not available through formal testing. It presents the opportunity to see the trial and error attempts students make in imperfect practice:

> Observing this experimentation, whether it occurs as children are playing or engaging in shared reading and writing activities with caregivers or teachers, affords an opportunity to document their understandings about literacy. . . . Their attempts at reading aloud, whether during pretend reading or as they follow print during shared, guided, or independent reading, reveal their understanding of the connections between sounds and symbols, as well as their knowledge of how print connects to meaning. (Casbergue, 2010, p. 18)

Carbonneau and colleagues (2019) collected observational scores of behavior, such as "Child initiates contact and play with other children and takes leadership roles by organizing activities and teaching peers." The reliability of those scores (230 observations of 23 three- and four-year-olds) was good, with next-to-same inter-rater agreement ranging from 84 percent to 95 percent over ten dimensions. Data collection used a platform called inCLASS that is widely used for observational scoring and has a research base (the inCLASS website lists 19 recent publications on data using the system).

A meta-analysis of the accuracy of teacher observations can be found in Südkamp et al. (2012). The type of rating that most closely corresponds to the DASL ratings is where teachers are asked to assess grade level equivalence for students. Südkamp et al. report an average correlation of .7 between such ratings and objective criteria for such studies. In one large-scale study, Meissel et al. (2017) caution about rater bias, especially in the case of high-stakes assessment.

Our conclusions from the teacher observation studies are that such assessment methods have a good research foundation and that observational rating is reasonable in low-stakes assessment of college learners.

2.3 The Evolution of Writing Assessment

Behizadeh and Engelhard (2011) give a historical perspective of writing assessment, emphasizing measurement theories, writing theories, and their expression in assessments of writing. A summary table outlines research traditions from 1900 to 2010, tracing a measurement focus from testing to "rater-mediated assessments," (p. 192) and an evolution of writing theory from writing as skills (error analysis) to writing in a sociocultural context, culminating with portfolio reviews. The authors cite the advent of portfolio reviews as a notable integration of theory into assessment:

> [W]e did not find significant evidence that writing theory influenced writing assessments. This is one of the major findings of our research and is of great concern. The one major exception to this rule occurred in the 1990s and continues into the present with writing theorists and teachers pushing for alternatives to multiple-choice and standardized essay tests, resulting in alternative assessment schemes such as portfolios. (Behizadeh & Engelhard, 2011, p. 205)

We see observational assessment as a natural extension of portfolio reviews and one potential answer to the authors' call for "innovations in classroom assessments of writing" (Behizadeh & Engelhard, 2011, p. 207).

The development of measurement theory in the Behizadeh and Engelhard (2011) account ends with item response theory and Rasch measures. These are hierarchical methods, which can usefully be extended to include explanatory variables that describe individual students. The motivation stems from the desire to build better regression models and to detect inequities in education that are linked to student characteristics.

The extension of portfolio ratings to observational (that is, fully contextualized) ratings and the extension of item response methods to richer hierarchical models addresses Behizadeh and Engelhard's concern that writing theory is not informing assessment. The inclusion of explanatory variables allows us to measure the connection between time and ability and estimate the impact of individual student characteristics like race, gender, and grade averages. This freedom of analysis leads to rich, testable hypotheses that simultaneously inform theory (through confirmation or falsification) and assessment (through error analysis).

3.0 Research Questions

One goal of the DASL project is to increase our understanding of how college students learn to write over four years, with the dual purpose of informing academic departments about student averages for normal assessment work and the broader research project that this paper comprises,

for which institutional review board approval was obtained before data-gathering began in 2015. The primary research question is

> RQ1. Can we plausibly model college average writing development over four years using the ratings we have collected?

In other words, are the instructor ratings useful in the aggregate, or are we just collecting random numbers? There are reasons to be dubious about the data. But if the ratings do give useful information, the collection method presents universities with a new low-cost source of data on student writing.

The answer to RQ1 entails skeptical inquiry into the reliability and validity of the data as well as technical challenges in summarizing (modeling) the results. Assuming that we find a useful model, then

> RQ2. How does writing development vary by student characteristics?

The second research question supports the first as another kind of validity check in that we expect sensible patterns to emerge.

These research goals are translated into claims supported by validity arguments found in the discussion section.

4.0 Research Methodology

4.1 Ratings Data Collection

The data come from ratings of students' academic performance over a 15-week college term rather than from decontextualized work samples. Course instructors were invited via an email link to an online survey to rate student learning outcomes using a format that looks like a grade sheet, except that instead of assigning grades, a rubric scale is used to assign developmental levels to learning outcomes. The invitation to participate in the survey was emailed to instructors two to three weeks before the end of the fall or spring term and remained open for about a month afterward. Each rating was stored in a database along with the name of the outcome being rated, student ID, rater ID, course ID, and other information. These ratings were scrubbed for data issues like accidental duplication of records, although these were few.

Ratings were gathered each fall and spring term from Fall 2015 through Fall 2019 by asking course instructors to voluntarily participate. Writing was one of about 100 learning outcomes that instructors could be asked to rate. The other learning outcomes used in this report were filtered to those with at least 1,000 samples. Instructors were asked to rate up to five student learning outcomes per course, but only when they believed they had a basis for judgment. Otherwise, the instructions were to leave the rating blank. Selections for the level of achievement were made using drop-down menus for each student, which defaulted to a blank (no rating).

The rubric is a generic developmental one that can be applied to learning outcomes that are cumulative over the four years of a typical bachelor's degree. The intent is to rely on the

faculty's professional training and experience rather than the detailed definitions of an analytic rubric. The developmental rubric was created in collaboration with faculty members from across disciplines and adjusted over a number of years to reach its present form. The model behind the design is Astin's (1984, 1996) interaction between student involvement, personal characteristics, and the environment in producing outcomes. Here, involvement entails all the aspects of writing that are encountered in course work. Because most academic disciplines at Furman University use writing as a primary tool for learning and assessment, we would expect that the quantity of involvement accumulates over time, for example, in numbers of papers written or time spent on writing processes.

Table 1

Rubric for Assigning Levels of Proficiency for Developmental Learning Outcomes (Those That Are Expected to Increase Over Time)

Scale level	Description
1	The student is not demonstrating the skill or knowledge sophistication that aligns with the university's expectation of entering college students.
2	The student demonstrates a level of aptitude that we would normally expect of capable students who are first learning the discipline. They may confuse terminology, misapply concept or technique, or lack a solid base of knowledge, but they show the capability to progress.
3	The student demonstrates the beginnings of sophistication with the discipline, making fewer elementary mistakes (terminology, concepts, techniques), and beginning to adapt to the culture (language, use of standard techniques, etc.).
4	The student demonstrates solid competency in the discipline's core knowledge and skill areas. He or she shows examples of synthesizing information, concepts, and application of techniques across the boundaries of individual courses and makes few elementary mistakes.
5	The student demonstrates the level of competency we expect of our graduates.

Note. This rubric is used to assess writing in the discipline, oral communication, and other outcomes. The prompt for discipline writing is "Writing effectiveness in the style most appropriate to the course, including knowledge of conventions."

The developmental levels in Table 1 are anchored by the ends of the scale. The two lowest levels distinguish between students who, in the instructor's opinion, are not yet doing college-level work (level 1) and those who meet the threshold expected for entering students (level 2). The other end of the scale is tied to expectations that instructors have of college graduates. The two intermediate levels (3 and 4) are described more generally to give instructors the latitude to interpolate between the endpoints.

This developmental rubric contrasts with the stricter definitions of an analytic rubric. The trade-off is between the appearance of precision with the analytic style and the use of natural language (e.g., "ready to graduate") in the developmental version. There are limitations to the

latter in that it assumes that the rater has enough experience to reasonably judge the upper and lower limits of outcomes. At a liberal arts university with small classes and much student-professor interaction—where these ratings were gathered—the conditions are good for success. We expect that raters are making holistic subjective judgments based on their experience as teachers and scholars. Their teaching styles and marking of papers vary greatly, so the ratings are nothing like standardized scores. We accept the trade-off in higher variability to gain greater contextualization and large sample sizes.

The learning outcome of main interest here is called "Discipline Writing," which is described as "Writing effectiveness in the style most appropriate to the course, including knowledge of conventions." Faculty-assigned DASL writing ratings were obtained from 67 to 84 percent of the 2012 to 2014 entering cohorts of undergraduates, 91 to 98 percent of 2015 to 2018 cohorts, and 59 percent of the 2019 cohort. Of this sample, 20 percent reported a race other than White, 15 percent were Pell grant eligible, 11 percent reported being first-generation college students, 10 percent were student-athletes (NCAA Division II), and 58 percent were female. These percentages are representative of the undergraduate population.

Foreign language skills are rated in most foreign language classes, including those taken for general education requirements. Music skills are only assessed within the music department. Critical thinking, oral communication, and writing are assessed in most disciplines and at all course levels. All of these are rated on a 1-5 scale designed to assess a student's development over time.

4.2 Explanatory Variables

Following Astin (1984, 1996), student learning is likely to depend on other factors than time in school. For example, Sullivan and McConnell (2018) found no plausible pattern of improvement in rubric ratings of writing samples from college freshmen to seniors unless assignment difficulty was taken into account. Accordingly, a variety of inputs are provided to regression models, including the following:

- Term, a student's time in college, measured in (fall/spring) terms 1-8, covering four years. This is the basic involvement index, assuming time in school means more development in writing.
- HSGPA, a student's grades in high school. These are standardized and recalculated from transcripts during admission so that they are on a consistent 4.0 scale.
- A student's college grades. We tested three types of these: first-year grade average (FYGPA), cumulative grade average when the rating was made (cumGPA), and a relative cumulative average over all grades (rGPA).
- A student's survey response to questions that ask about self-confidence in writing and other academic skills. The survey was designed by the Higher Education

Research Institute and is administered nationally each year. The survey is administered to first-year students early in the fall term, with response rates of about 80 percent.

- StudentID, a unique student identifier used for the hierarchical models.
- RaterID, a rater identification included in the data collection of ratings, so that rater leniency or severity can be accounted for.
- Subject, the subject of the course where a student received a writing rating.

The information is housed in databases and retrieved through automated scripts. Everything is checked for duplicates, missing values, and other data problems that might arise. No imputation was done for this study; missing values caused samples to be ignored.

4.3 Reliability

Reliability is the foundation of validity in the sense that we want some assurance that the ratings are not merely random numbers: is there enough "signal" to discern useful trends beyond the noise? A subset of this dataset was the subject of an initial study of inter-rater agreement, found in Eubanks (2017), using a kappa statistic.

Because the intent of the current work is to analyze trends within the data, the kappa-type inter-rater type of reliability is less useful than variance-based methods. We think of the ratings as imperfect measures and estimate residual error within and between groups. This approach is common in hierarchical models and is related to generalizability theory and analysis of variance. An overall measure of reliability is the intraclass correlation (ICC) defined in Shrout and Fleiss (1979) that compares within-student variance to between-student variance. The former is seen as error, and the latter is the ability to distinguish cases. The ICCs given in this paper are for single measures: the correlation of two random samples under the same conditions. The ICC increases when using averages instead of individual ratings according to the Spearman-Brown formula.

4.4 Validity

The suitability of DASL ratings to describe student writing development is explored through comparisons with other types of data, using regression models and factor analysis. Comparisons and contrasts correspond to the established analyses of convergent and discriminant validity and extend to the interpretation of constructs induced through factor analysis. We use a varimax (non-orthogonal) rotation (Kaiser, 1958) as a compromise between dimension reduction and ease of interpretation. It was performed using the R programming language's base function for that purpose.

4.5 Regression Methods

To find patterns in the data, we built regression models that assume that the ratings are scalar and then checked the results with ordinal methods. Scalar models estimate score averages, which will

be familiar to many readers. The ordinal models estimate probabilities of a rating reaching a threshold.

The models used are hierarchical, sometimes called mixed effects models, and are similar to item response theory (including Rasch methods), partitioning variance into raters and subjects. This approach is sometimes justified by assuming that there is a latent (or invariant) scale that is not observed, but by inducing it we can simultaneously assess student ability and rater severity with a common metric (Engelhard & Wind, 2017). Statistically, hierarchical methods are attractive because they allow us to make use of all the data, even when students have varying numbers of ratings. The models assume that the random effects (e.g., coefficients for each student and rater and subject) are normally distributed and effect "shrinkage" on these estimates to increase the precision in estimating fixed effects. The latter coefficients describe average development over time and are of primary interest.

The models assume that students have individual ability levels that can't fully be explained by averages or probabilities. Raters may be more or less severe, and ratings within a subject like chemistry might be different from ones in English. But after subtracting out these sources of variation, we are left with a simple idea: average development over time is a line that can start higher or lower depending on a student's grade average and have a slope (rate of increase) that also depends on grade averages.

The regression analyses were performed using the R programming language augmented by the *lme4* and *ordinal* packages, which are designed to solve mixed effects problems.

5.0 Results

5.1 DASL Rating Overview

From Fall 2015 through Fall 2019, course instructors have contributed 167,606 ratings of student learning outcomes. Of these, 18,536 are writing ratings on a scale intended to show change over time, assigned by 289 unique course instructors and covering 4,774 undergraduates, or about four ratings per student. The 2015 entering cohort was assessed each of the eight fall and spring terms to graduation (the four-year graduation rate for that cohort was 77 percent).

We will contrast the writing ratings with other types of learning outcomes ratings present in the data set. The counts and proportions of ratings for these are shown in Table 2.

Table 2

Counts and Distributions for the DASL Learning Outcomes Used

	Rubric rating				
Learning outcome	**1**	**2**	**3**	**4**	**5**
For. lang. listening (1,904)	0.03	0.34	0.33	0.28	0.02
For. lang. speaking (2,081)	0.06	0.36	0.31	0.25	0.02
For. lang writing (2,091)	0.05	0.35	0.35	0.23	0.02
Music performance (1,074)	0.02	0.10	0.48	0.27	0.13
Music technique (1,038)	0.02	0.15	0.45	0.27	0.12
Oral comm. (3,142)	0.02	0.14	0.36	0.26	0.21
Critical thinking (10,223)	0.03	0.14	0.42	0.25	0.16
Discipline writing (14,093)	0.02	0.14	0.44	0.25	0.15

Note. Each row gives the proportion for responses 1 (lowest) to 5 (highest), summing to one. For example, oral communication has the highest proportion of 5-ratings among these outcomes (21% of ratings). The discipline writing outcome has a distribution centered at 3.

The distributions of ratings in Table 2 are not as skewed (right-censored) as course grades, and the least-skewed outcomes are those used for foreign language courses, where level 5 is aspirational, rather than an expectation for graduates. Even for the common developmental scale with 5 = "ready to graduate," the raters are conservative.

A factor analysis of these ratings was performed by correlating contemporaneous ratings per student and then using a varimax (non-orthogonal) rotation. Three factors are sufficient to capture 82 percent of the variance (33%, 25%, and 23%, respectively).

Table 3

Factor Analysis Loadings Using Varimax Rotations for the DASL Learning Outcome Ratings (1-5) With the Largest Sample Sizes

	For. lang.	Academic	Music
For. lang. listening	0.90		
For. lang. speaking	0.94		
For. lang writing	0.89		
Music performance			0.96
Music technique			0.87
Oral comm.		0.65	
Critical thinking		0.76	0.31
Discipline writing	0.33	0.92	

Note. Loadings less than 0.3 are suppressed for readability. The column headers are creations intended to be suggestive of the groupings.

The loadings in Table 3 neatly cluster the DASL outcomes into a foreign language proficiency, academic thought and communication, and musical proficiency, showing that faculty raters are distinguishing between these learning outcomes. The DASL writing ratings are linked to oral communication and academic thinking skills, an indication that the writing ratings represent general skills in addition to discipline-specific ones (the prompt asks for "writing in the discipline").

5.2 Writing Ratings

To examine the writing ratings in more detail, we show raw counts by the term, where 1 = the first fall and 8 = spring of the fourth year. Some students have more than one rating in a term.

Table 4

Counts of Raw Writing Ratings by Rating Assigned and Term Number When the Rating Was Assigned

Rating	Term							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	132	98	72	34	37	47	39	50
2	956	598	501	307	274	173	152	137
3	1,058	1,052	1,003	1,110	775	844	691	542
4	428	390	460	660	683	769	884	711
5	44	49	85	134	246	386	783	1,149

Note. Ratings are on a scale of 1-5, with 1 being the lowest. Term 1 is the first fall term of the freshman year, and term 8 is the final spring term of the senior year.

The rating counts in Table 4 in the first two terms show some inflation beyond the intended purpose of the rubric, where a rating of two means acceptable first-year work. Somewhat more students received a 3-rating than a 2-rating their first term. There is an evident trend toward higher ratings over time, which we will model below. The density of data across ratings and terms is appropriate for the study. The low cell counts occur where we would expect them: in the lowest and highest categories.

To estimate reliability, cases were filtered to those where the same student was rated at least twice in the same term, giving a 95 percent confidence interval ICC = .45, 95% CI[.44, .45] with 10,371 ratings of 4,649 students. When the data were further restricted so that ratings were within the same term and same academic discipline, the confidence interval was ICC = .49, 95% CI [.46, .52] with 3,428 ratings of 1,581 students. When the data were grouped by year in college, the general ICC estimates were .23, .29, .31, and .33 for years one through four, respectively (these are smaller because of the more limited range of measured abilities). More variance statistics are found in Table 6.

5.3 Grades

Course grade averages are used as a proxy for general academic ability in predicting writing ratings. We assume that grades are influenced by a student's intelligence, prior education, work habits, and other personal factors that we cannot measure individually. Grades are also affected by the courses taken, how they are taught, and how grades are assigned. At our university, grade reliability varies by discipline, from around ICC = .2 for the arts to ICC = .7 for foreign languages. These numbers are similar to those found by a large multi-institutional study (Beatty et al., 2015), with ICCs from .32 to .56, depending on the discipline.

For regression analyses, we define a relative grade average (rGPA) that accounts for course difficulty. The rGPA distribution is approximately centered at zero, and a score of 1 means the student, on average, earns a grade point higher than the class average. Very few students

consistently earn a grade point higher than the class average, as can be seen at the far right of Figure 1B. This distribution is much nicer statistically than raw grade averages, which bunch up near 4.0 (Figure 1A).

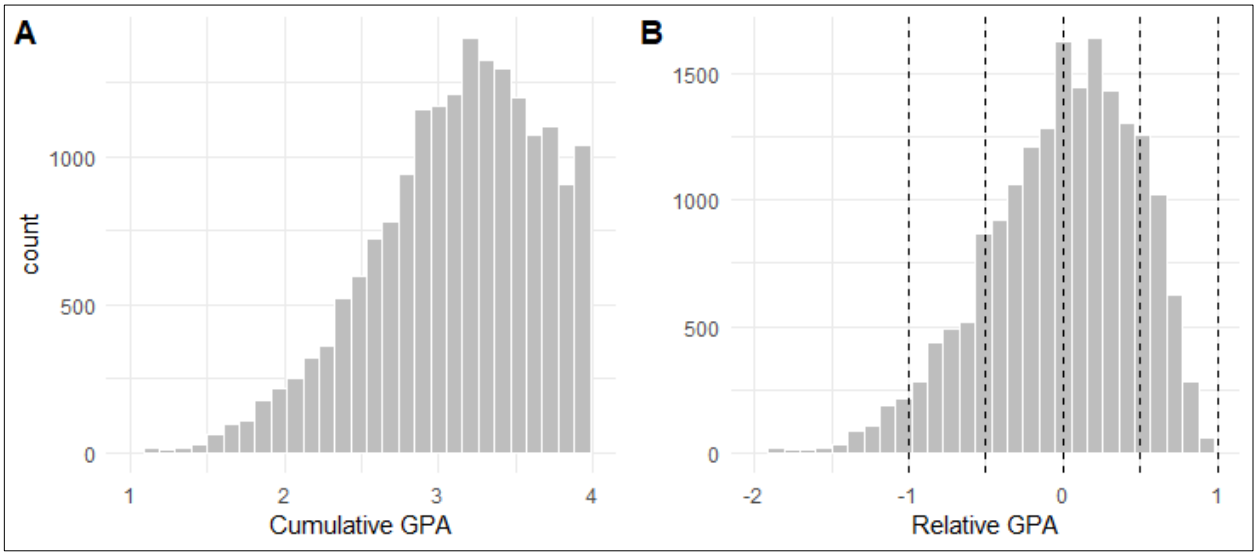

Figure 1

Histograms of College Grade Averages Used as an Explanatory Variable in Regression Models

Note. *Plot A shows cumulative grade averages as of the time the respective DASL rating was made. Plot B shows relative grades, calculated by subtracting the course grade average from each student's grade in that course and then averaging those over each course the student has taken. A relative grade of -1 means a student who on average earns a grade point lower than the class average.*

As a precaution against grades having too much influence over writing ratings, we filtered out all the grades that came from courses where a student received a rating. However, this made no difference to the results.

Because cumulative GPAs are affected by survivorship bias (only students who made it to their junior year have grades that year, etc.), it is important to check for longitudinal trends that might affect the writing development models. A visual check of each of the graphs of raw GPAs and relative GPAs showed that there are no consistent term-based trends in either measure that would bias the models. In other words, grade averages tend to be stable over time.

In predicting writing ratings, we used high school grades (HSGPA), first-year college grades (FYGPA), cumulative GPA at the point the rating was assigned (cumGPA), and the relative average just described (rGPA) using all available grades. Each of these predictors was used to model writing ratings with

$$Rating_i = \beta_0 + \beta_1 Time + \beta_2 GPA + \beta_3 GPA * Time + \gamma_{S_i} + \gamma_{R_i} + \varepsilon_i,$$

which is the form used throughout this report. As will be explained in more detail below, the primary interest is in β_1(growth rate), β_2(impact of grades), and β_3 (grades' effect on growth rate). The resulting coefficients appear in Figure 2.

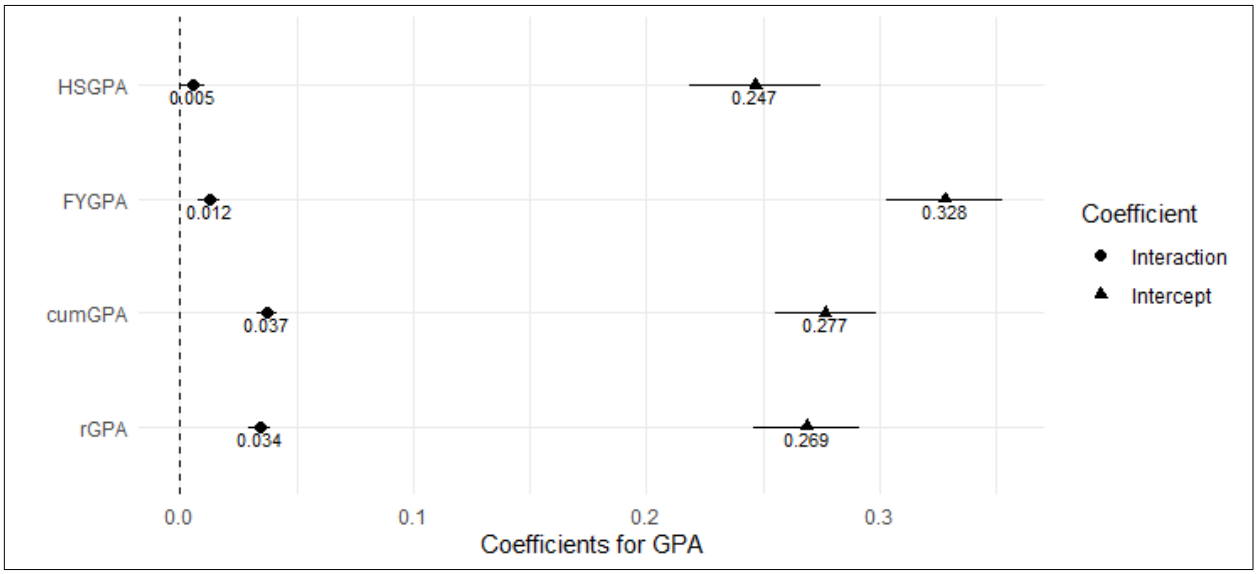

Figure 2

Regression Coefficients Predicting DASL Writing Ratings, Showing the β_2 Coefficient for GPA and the β_3 Interaction Term Between GPA and Time

Note. *HSGPA is a recalculated high school grade average (R^2 = .23), FYGPA is first-year college grade average (R^2 = .30), cumGPA is the cumulative GPA when the rating was assigned (R^2 = .34), and rGPA is the relative GPA over all available grades except those a student received a rating in (R^2 = .32). The GPA measures were first scaled to have mean zero and standard deviation one so that the coefficients would be comparable. Lines denote 95 percent confidence intervals.*

The evident pattern in Figure 2 is that as information about college grades increases, the interaction term increases as well. Cumulative grades and relative grades give about the same answer. Positive values for the interaction term imply rates of development that depend on grade averages.

Of the models, cumGPA and rGPA had the best model fit (R^2 = .34 and .32, respectively). Because rGPA has better statistical properties and because it is more likely to be comparable to studies at other institutions in the future, we chose to use rGPA in the final models.

5.4 Self-Ratings of Abilities and Traits

The DASL ratings come from professors assessing their students. Students also rate themselves on surveys, and we would expect some convergence between the two perspectives. The source is a first-year survey from the Higher Education Research Institute (HERI), called The Freshman Survey, which is administered each fall to the entering class of undergraduates. One group of

items asks each student to compare his or her abilities to those of peers, with possible responses being 1 = Lowest 10%, 2 = Below Average, 3 = Average, 4 = Above Average, and 5 = Highest 10%. Since writing ability is one of the self-ratings, it creates a natural test for comparison to the DASL ratings. There are 3,121 complete cases (data on each self-rating) that also have student IDs, with 12,006 matched DASL ratings.

Table 5

Factor Analysis Loading Using Varimax Rotations for Freshman Survey Self-Ratings

Self-Rating	Creativity	Leadership	Academics	Health	Writing
Artistic ability	0.99				
Creativity	0.54				0.31
Leadership ability		0.96			
Public speaking		0.44			
Mathematical ability			0.51		
Academic ability			0.91		0.32
Physical health				0.54	
Spirituality				0.33	
Drive to achieve				0.31	
Emotional health				0.65	
Writing ability					0.66
Understanding of others					

Note. Self-ratings are on a 1-5 scale, with 5 being the highest (N = 3,121 complete cases). Loadings less than 0.3 are suppressed for readability. The column headers are creations intended to be suggestive of the groupings.

To assess the internal validity of the self-ratings, a factor analysis was performed on the paired item correlation matrix with varimax rotations. The results in Table 5 show that five factors can account for about half the variance (ranging from .11 to .07 left to right on the table). The loadings have straightforward interpretations, which suggested the labels we chose for the factors. Writing is most associated with academic ability and creativity. This provides some evidence that the self-ratings are a useful comparison to DASL ratings.

Based on the factor analysis results, we suspect that when new first-year students rate their writing ability, they tend to think more of creative writing than academic writing, so we created another version of the writing rating (writing ability*) by regressing it against creativity and taking the residuals as a new measure of writing minus creativity.

We tested the association with DASL writing ratings by correlating student self-assessments with first-year and fourth-year DASL averages. We did the same for grade averages to add context. The result appears in Figure 3.

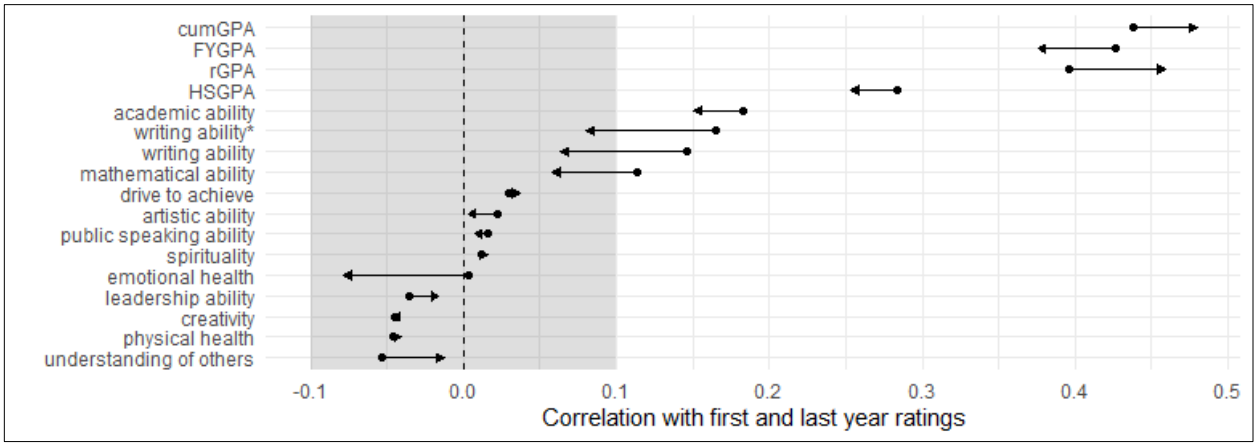

Figure 3

Pearson Correlations Between DASL Writing Ratings and Grades (GPA Measures) or Self-Ratings on a Freshman Survey

Note. *The line segments extend from the correlation coefficient in the first year (dot) to the coefficient in the fourth year (arrowhead). The starred writing ability is a construct that regresses out the influence of the creativity rating. Values near zero are shaded to indicate coefficients so small as to be inconsequential.*

Since the self-ratings are gathered at the beginning of the first year of college, we would expect that any explanatory power they have over instructors' writing ratings would wane over the four years, meaning that the arrows in Figure 3 would point toward zero (the vertical dashed line). This pattern is evident in most of the measures, but the predictive power of the academic ability self-rating is particularly impressive three years after the survey. The writing self-rating without creativity performs marginally better than the original, but both wane in importance by the fourth year. All the grade measures perform better than the self-ratings, with college grades better than high school grades.

If instead of the DASL writing outcome, we correlate with DASL musical performance or technique, only the mathematics self-rating shows a correlation to the ratings comparable to writing or academic ability in Figure 3 (and only in the first year; it drops to zero in the fourth year). Grades are still predictive, but less so, with correlations of around 0.3 instead of around 0.4.

The pattern of correlations between student self-assessments and DASL instructor ratings support the latter as a general construct measuring writing.

5.5 Modeling Ratings as Scalars

It is a common approximation with rating scales to assume that they are scalar, implying that the gap between a 1 and 2 is the same in measurement as the gap between 4 and 5. In some cases, this assumption may not be valid, but because of the ease in interpreting results, we start with averages.

A linear regression model was created that includes random effects (individual intercepts but not slopes) for each student, rater, and subject of study (from the course designation, e.g., BIO 101). These coefficients "subtract out" the effects of biases from varying numbers of students, raters, and subjects, accounting for rater severity, for example (Agresti, 2013; Gelman & Hill, 2006). Mathematically, the scalar model is

$$Rating_i = \beta_0 + \beta_1 Term + \beta_2 rGPA + \beta_3 Term * rGPA + \gamma_{S_i} + \gamma_{R_i} + \gamma_{D_i} + \varepsilon_i,$$

where *i* indexes each rating, and the β coefficients are the fixed effects for a constant (β_0), the term number (1-8) when the rating was made (β_1), the relative grade average (β_2), and the interaction between the term and relative grade average (β_3). The indexed γs are random effect coefficients assigned to each student (S_i), rater (R_i), and discipline of study (D_i). The last term (ε_i) captures residual error for the *i*th rating.

Table 6

Regression Summary for the Mixed Effects Model Assuming Scalar Rating Values

	Rating		
Predictors	***Estimates***	***95% CI***	***p***
(Intercept)	2.729	2.629 – 2.828	<0.001
Term	0.160	0.154 – 0.165	<0.001
rGPA	0.545	0.500 – 0.590	<0.001
Term * rGPA	0.068	0.059 – 0.077	<0.001
Random effects			
σ^2	0.44		
$\tau_{00\ StudentID}$	0.05		
$\tau_{00\ RaterID}$	0.20		
$\tau_{00\ Subject}$	0.07		
$N_{StudentID}$	4,770		
$N_{RaterID}$	289		
$N_{Subject}$	51		
Observations	18,537		
Marginal R^2 / Conditional R^2	0.297 / 0.593		

Note. Term is the term number (1-8) of the student's enrollment when the rating was made, and rGPA is the relative grade average over all available grades. Random effects are per student, rater, and course subject.

The fixed effects at the top of Table 6 have convincingly non-zero coefficients. The Term coefficient represents the average ratings gain from a term of instruction. Ideally, students start at a rating of 2 and graduate with a rating of 5 after seven terms of growth, an average of .43 per term. The actual value is about a third of that because of scale compression and stratification associated with grade averages. A one-letter average grade advantage (rGPA = 1) translates into an additional .53 rating points, more than three terms worth of development. The growth rate for those students also increases by an additional .07 rating point per term, or .49 over seven terms, for a total of about 1 rating point bonus at graduation.

The random effects section of Table 6 serves as an analysis of variance or reliability analysis. It shows that allowing each rater a severity level accounts for a significant amount of variance, probably because each rater is responsible for many ratings on average. Student and subject random effects account for less variance, but we chose to leave them in the model for illustration. Adding random slopes to each student does not help the model.

The two R^2 values given in Table 6 follow Nakagawa and Schielzeth (2013). The marginal value assumes that the random effects are not counted in explaining model fit, whereas the conditional value assumes they are. About half of the model's effectiveness in fitting the data is due to the inclusion of the random effects, most of which are from rater offsets.

To visualize model fit, we plotted conditional rating averages over time and compare those to the fitted values from the model. Each rating was binned into a quintile by rGPA, with 1 = the lowest 20 percent and 5 = the highest 20 percent.

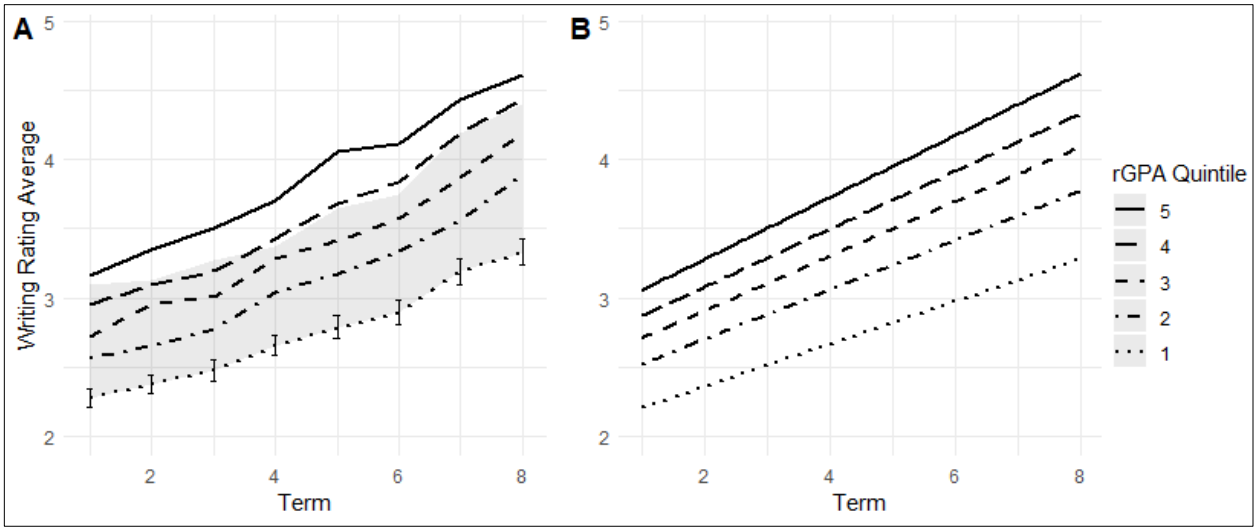

Figure 4

Student Average Ratings Over Time Are Grouped by Term Number and Quintile of Relative Grade Averages (rGPA)

Note. Plot A shows empirical averages with the first quintile annotated with vertical lines showing a 95 percent confidence interval and a gray region that illustrates the size of the standard deviation of the ratings. Plot B shows the modeled averages smoothed over the varying characteristics of students using the model in Table 6.

The shading in Figure 4A shows that the standard deviations for ratings in the first rGPA quintile are quite large. Consequently, the ratings should not be used to guide students as individuals. However, the confidence intervals on the averages are reasonable, even separating the rGPA quintile curves. We conclude that the model is a reasonable indicator of the average change over time.

The effect of the positive interaction term between Term and rGPA is a divergence of growth paths, resulting in the fan shape in Figure 4. The main features of the average ratings are captured in the model, including a larger gap between the first and second rGPA quintiles, but there is a quadratic curve to the lower quintiles that the linear model does not represent well. Squared terms for the time and rGPA terms help the model fit, but at the cost of too much complexity for our purposes here.

The empirical averages and the model support the claim that instructors give higher ratings over time on average but that growth rates depend on grades.

5.6 Modeling Ratings as Ordinals

Because the DASL writing ratings are ordinal, the analysis that assumes they are scalars violates some of the model assumptions (e.g., residuals are not normally distributed). To check the result, we replicated the regression model using methods designed for ordinals. The "proportional odds" model is

$$\ln\left(\frac{\Pr\left[Rating_i \leq k\right]}{\Pr\left[Rating_i > k\right]}\right) = \tau_k - \beta_1 Term - \beta_2 rGPA - \beta_3 Term * rGPA - \gamma_{S_i} - \gamma_{R_i} - \gamma_{D_i} - \varepsilon_i,$$

where the fixed and random effects are as before, but the constant β_0 is replaced by breakpoints $\tau_k, k = 1,..,4$. The model is still linear, but rather than in rating units, it is linear in log-odds (the left side of the equation). The coefficients are subtracted in this model so that a positive term conceptually shifts the τ_k to the left, making it more likely that a threshold will be passed.

Table 7

Regression Summary for the Mixed Effects Model Assuming Ordinal Rating Values

	Rating		
Predictors	***Lzog-Odds***	***CI***	***p***
1|2	-3.801	-4.123 – -3.480	<0.001
2|3	-0.717	-1.021 – -0.412	<0.001
3|4	2.333	2.027 – 2.640	<0.001
4|5	4.797	4.482 – 5.112	<0.001
Term	0.472	0.453 – 0.491	<0.001
rGPA	1.571	1.432 – 1.710	<0.001
Term * rGPA	0.222	0.193 – 0.250	<0.001
Random effects			
σ^2	3.29		
$\tau_{00\ StudentID}$	0.46		
$\tau_{00\ RaterID}$	2.03		
$\tau_{00\ Subject}$	0.60		
$N_{StudentID}$	4770		
$N_{RaterID}$	289		
$N_{Subject}$	51		

Note. τ_1 *is denoted 1|2, the cut point between the first and second rating level, etc. Term is the term number (1-8) of the student's enrollment when the rating was made, and rGPA is the relative grade average over all available grades. Random effects are per student, rater, and course subject.*

The coefficients generated from the proportional odds model are found in Table 7. They cannot be compared to those in Table 6, but their signs have the same meaning. Specifically, the influence of all three fixed effects (Term, rGPA, and their interaction) are positive: increasing values predict increased probabilities of higher ratings.

The proportional odds model predicts probabilities of ratings after making the simplifying assumption that the coefficients are fixed for each rating cut point τ_k. Model fit is assessed by

comparing the proportional odds model to individual logistic regressions, where $\beta_1, \beta_2, \beta_3$ can vary by cut point instead of being fixed for all cut points. In those individual models, the coefficients generally agreed with the proportional odds model, but the interaction term β_3 was convincingly non-zero only for the first two models (Rating > 1 and Rating > 2). For that reason, the plotted comparisons in Figure 5 use the individual logistic models for slightly better accuracy.

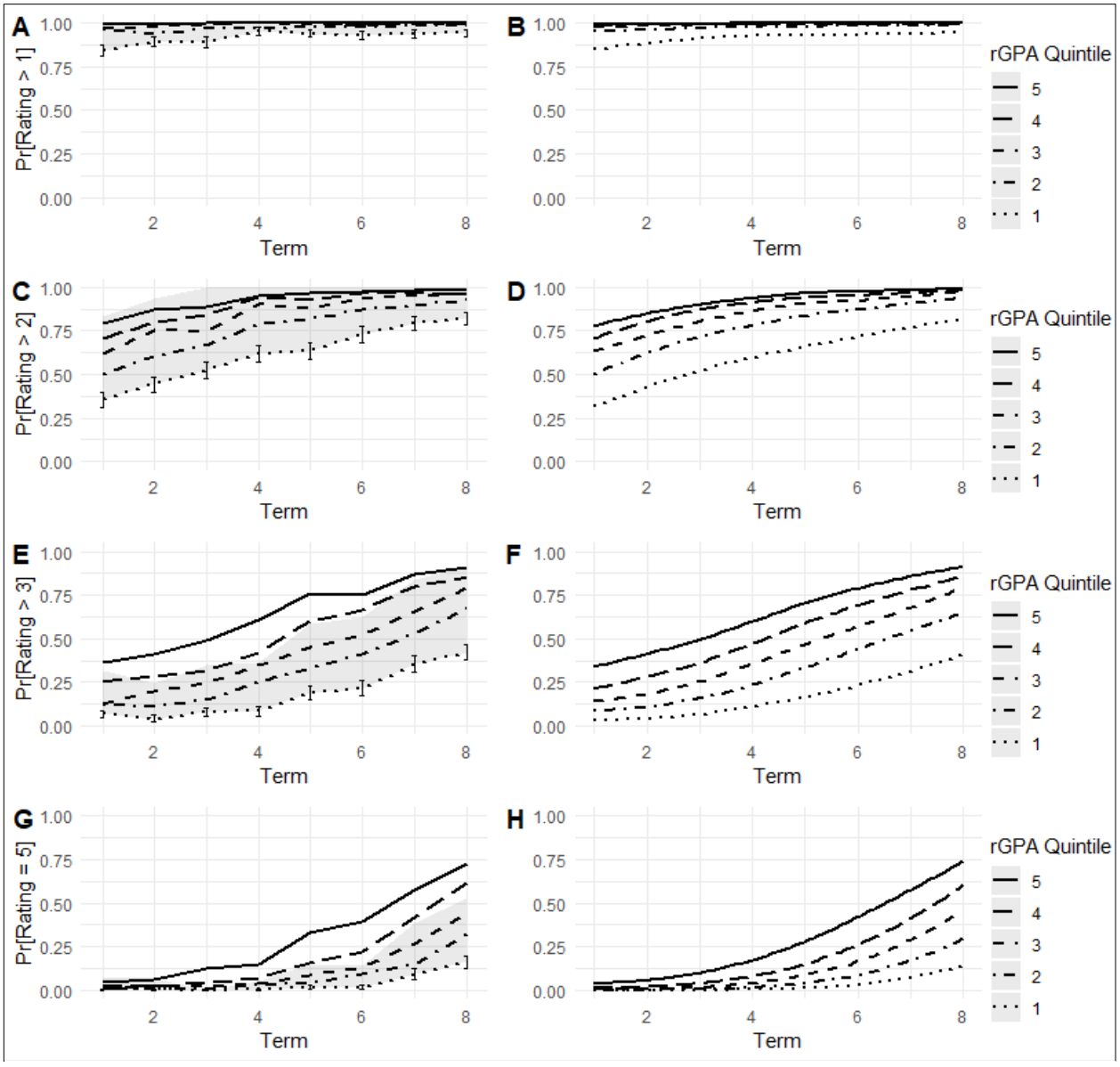

Figure 5

Empirical Proportions (Left) and Estimated Probabilities (Right) of Writing Ratings Greater Than Thresholds 1 (A, B), 2 (C, D), 3 (E, F), or 4 (G, H) Over Time (Term) and Quintile of Relative Grade Averages (rGPA). See also the Note *on the following page.*

Note to Figure 5. Left plots show empirical proportions with the first quintile annotated with a 95% confidence interval and a gray region that illustrates the size of the standard deviation of the probability. Ranges for the other quintiles are suppressed for readability. Right plots show the modeled probabilities for each of three corresponding logistic regressions. The coefficients in Table 7 correspond to an average model for B, D, F, and H.

The lowest rGPA quintile is gapped in the probability estimates (Figure 5B, D, F, and H), and the fan shape is particularly evident in plot H. An average student (quintile 3) has about a 40 percent chance of receiving the highest writing rating in term 8, while a student in the next lowest group only has about a 30 percent chance.

The B plot tells us that even the lowest quintile of rGPA has about an 80 percent chance in term 1 of attaining the rating expected of new students (or better rating). However, the other students have a nearly 100 percent chance of doing so.

In the linear model, we saw evidence of rating inflation in the early terms, and the same can be seen in Figure 5C, where even the lowest rGPA quintile of first-term students has about a 40 percent chance of receiving a rating of three or higher. This finding aligns with the lower rater agreement for that end of the scale noted in Eubanks (2017).

The two types of regression models give us similar information. The visual fit of empirical data to fitted curves in both is reasonable, and the developmental growth in averages or probabilities is feasible.

5.7 Demographic Variables

As with any regression modeling, multiple variations were tested for fit, seeking a compromise between explanatory power and parsimony. In particular, several student characteristics in addition to rGPA were included in the linear model to test for importance: first generation college student status, Pell grant eligibility, race (White or non-White), gender, and (non-) athlete status. Of these, only race (non-White β = -.04, p = .02*)*, gender (male β = -.04, p = .004*)*, and athlete status (non-athlete β = .13, p <.001*)* were statistically significant. None of them improved model fit enough to include in the final model, but the statistics are included here since they relate to rating fairness across student types. The race and gender coefficients are too small to have practical significance, except possibly in combination: non-White males who are athletes are predicted to score .21 rating points lower than other students when all else is the same.

The demographic variables were included in a model that already included rGPA, so any effects of race, gender, and so on that cause higher or lower grades are already accounted for. In other words, unfairness related to a demographic group could be completely captured in how that group is assigned grades.

5.8 National Survey Data

A natural comparison to the faculty-assigned ratings was found in self-assigned writing ratings collected nationally by the Higher Education Research Institute (HERI, 2006), using the same

freshman survey described in Section 5.4, and here matched to a senior survey. The Institute releases old data for research purposes, which made available paired freshman-senior surveys from 1996 through 2006. The surveys come from four-year college completers, who answer over a hundred questions, many of which are matched in both surveys. These include the self-ratings that are featured in Figure 3 and Table 5. Here, we focus on the writing self-assessment responses from the freshman and senior surveys, combined with a self-reported grade average. The sample size is 237,314, with the grade responses from lowest to highest GPA having samples of 4,802, 25,005, 74,677, 96,403, and 35,955. This distribution approximately matches the distribution of grades used in the DASL models.

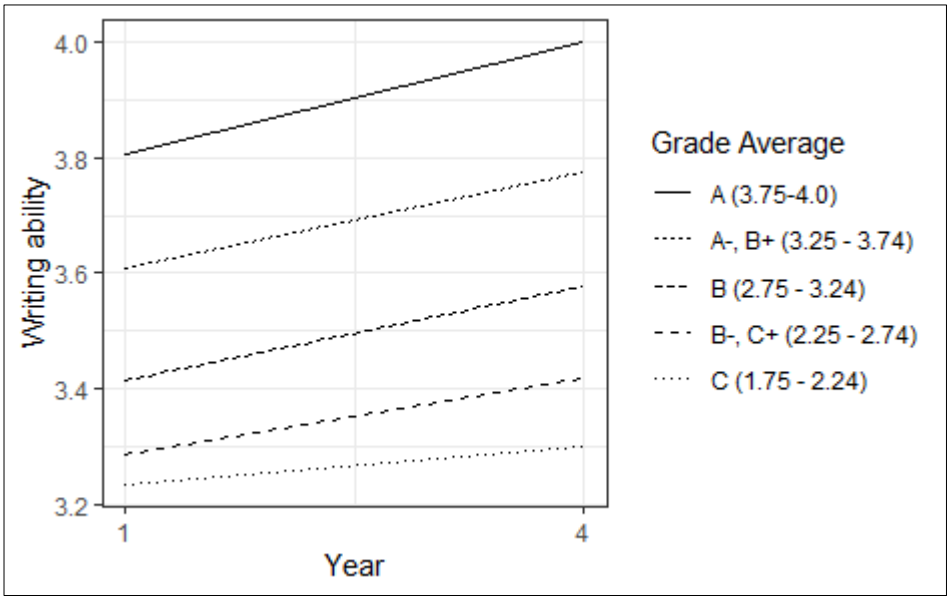

Figure 6

*Average National HERI Survey Self-Ratings (*N *= 246,000) of Writing Ability*

Note. Self-ratings are shown on a 1 (lowest 10%) to 5 (highest 10%) scale at years one and four, with connecting lines to show slopes. Averages are grouped by self-reported college grade averages given in year four.

To compare the national data to the faculty-generated DASL ratings, the survey responses were first grouped by the self-assigned grade average, then the first- and fourth-year self-ratings were averaged. The resulting divergence pattern appearing in Figure 6 is like the faculty rating averages summarized in Figure 4B.

The link between grades and writing self-efficacy has been studied before. For example, Martinez et al. (2011) used structural equation modeling on a survey of 127 college students to find that "Students with higher GPAs experience lower levels of writing anxiety than students with lower GPAs" (p. 356), and that higher anxiety led to lower self-efficacy. Pajares (2003)

undertook a literature review of writing self-efficacy and outcomes, finding that "Effect sizes between writing self-efficacy and writing outcomes in multiple regression and path analyses that control for preperformance assessments such as writing aptitude or previous achievement have ranged from .19 to.40" (p. 145).

Although the data are quite old, it is reasonable to generalize that students with lower self-assessments at entry to college improve these appraisals as they improve writing abilities. If the learning rates depend on grades, as suggested in Figure 4, it would explain the divergence pattern in Figure 6.

6.0 Discussion

The discussion follows the inferences in Kane's (1992, 2006) validity framework, centering on the two research questions. Based on the evidence, we claim that (1) the DASL writing rating averages describe development of a writing construct over time, and (2) on average, writers improve their skills over four years of college, but that (3) divergence patterns seen in Figures 4, 5, 6, 7, and 8B allow us to infer that college perpetuates or expands differences in writing skill levels.

6.1 Scoring

Claim 1: DASL ratings on average measure general writing ability.

It is useful to compare the DASL process to a portfolio review, where student work is assembled according to guidelines and then rated. The DASL ratings assume that course instructors who assign papers already process this kind of information, but over a fifteen-week semester instead of in a single scoring session. With portfolios, we have more control over the observations, for example, by specification of contents and rater training, while the DASL ratings are based on evidence that is unique to a course. Portfolio reviews emphasize the measurement task, while DASL ratings are fully contextualized within teaching and learning, and emphasize teacher judgment.

The DASL ratings' validity assumes that instructors only provide ratings when they have a basis for judgment, and that they apply the rubric intuitively, based on past teaching experience, to rank students against this historical comparison set, not just the current class. Such behavioral observation has a successful history in K-12 education, lending plausibility to the DASL exercise. The expertise of the faculty as judges must be acknowledged. At least in the context of the present study where the student-faculty ratio is 10:1 and there is a strong emphasis on teaching, it would be surprising if their observations had no value. Informal judgments made eventually by graduate advisors and employers presumably resemble the observational DASL ratings.

The ICC reliability estimates for individual DASL writing ratings are similar to course grades at the university (.49 for grades and .44 for the writing scores), and they compare well to rubric scores of portfolios in other studies. White et al. (2015) give a range of ICC from .21 to .55, depending on the trait being measured (Pearson correlation and ICC are the same with two

samples). Ross and LeGrand (2017) found ICCs from .19 to .3 in rating portfolio traits, and Kelly-Riley et al. (2016) found ICCs from .37 to .54 over six portfolio traits.

Reliability needs to be high when judgments have effects on individual students, but that is not the case with the DASL ratings, where we are only interested in trends. The complement of reliability (1 – ICC) is measurement error, and ICC in the linear regression becomes an R^2 value. In Table 6, the conditional R^2 of .593 is the fraction of the score variance captured by the model when the random effects are included. Those comprise nearly 5,000 additional coefficients (one each per student, rater, and subject). The fact that .07 score variance comes from differences between academic subjects is one indication of validity of the "discipline writing" intent. It means that student writing is judged differently across disciplines, as we would expect.

The reliability of the DASL writing ratings is too high to conclude that they are merely "random numbers." But do they actually measure writing? To address that question, we turn to the DASL ratings' relationship to other kinds of data.

The factor analysis of learning outcomes ratings (Table 3) shows that the writing rating correlates with thinking and communication skills but not with musical ability or foreign language proficiency. The DASL ratings apparently measure a general writing ability that correlates with academic performance. This should not be surprising, since earning high grades in many cases entails writing good papers. Kelly-Riley et al. (2016) found a correlation between holistic portfolio ratings and an introductory English course grade of .80, and correlation of the ratings with high school grade averages of .18. Comparative values for DASL ratings can be found in Figure 3, where the correlation of high school GPA with first-year DASL ratings is about .28.

The factor loadings for student self-assessments in Table 5 support their concept of writing as a construct entailing creativity, general academic ability, and writing as a specific ability. The correlations between student self-assessments in writing, mathematics, and so on illustrated in Figure 3 show a convincing link between the DASL writing scores and the self-assessments in academic ability, writing ability, and mathematical ability. The latter connection has been noted before, for example, in White et al. (2015). Our calculations from the HERI national survey data (Figure 6) provide another link between writing and grades (Figure 4) that agrees with the DASL averages (Figure 4). In the Furman data, ratings of music outcomes (technique and performance) do not show the same association with academic self-ratings.

Support for DASL ratings as measures of student writing comes from the convergence of similar measures, like student self-ratings and grades, and the differences in assessments of music or foreign language and differences from student self-ratings unrelated to writing, like spirituality or leadership ability. The regression results, showing positive growth over time, support the design of the rubric. The graphs in Figure 4 suggest that average writing abilities grow over four years of college, slightly accelerating in the later years. This is consistent with the concentration of an academic major's coursework in the last two years. There is, however, a threat to the validity of this conclusion.

Might raters just be using the course level (e.g., 300 versus 400) or the seniority of the student as a benchmark, and then adjusting the DASL rating up or down based on the student's grades? This would explain an increase over time in rating averages even when writing qualities are not changing, reducing the ratings to a proxy for grade averages.

One piece of evidence against this "social promotion" idea is that the highest rating (a 5 on the 1-5 scale) is described as the student meeting the expectations of a graduating senior. If raters were marking to expectations, then most seniors would receive a 5 at graduation, but in fact, the raters are rather stingy with the highest rating (Figure 5H), with only about half of graduating seniors achieving that distinction.

Additionally, if ratings were based on student seniority (freshman, sophomore, etc.) to a significant extent, we would expect the graph in Figures 4 and 5 to look more like stair steps, since students are generally "promoted" at the end of an academic year. In fact, the fall-spring increases on average are the same as the spring-fall increases.

If the development over time in writing ability is illusory, then faculty are mostly rating students on general abilities that are constant across four years of college, like grade averages. This implies that the faculty either are not teaching students to write, or that they are, but cannot perceive it. The details of how the ratings would have to work in that case are quite complicated. If raters are to produce the patterns in Figure 4 with a *pro forma* system, it must be a version of social promotion that is appropriately nudged up or down for academic ability while accounting for class level. For example, "Tatianna is a B- student who is a senior, so she probably deserves a rating of 3 in a 400-level class, unless it's a music rating, in which case we mostly ignore the grade." Since we removed the grades from courses where ratings were made, the link is to the grades that *other* instructors are assigning. Occam's razor suggests that at least some of the complexity is due to underlying states (i.e., latent variables). As noted above, the developmental curves for quintiles 1-4 in Figure 4A show a slight curve upward, which we would expect if discipline writing is learned more in the last two years than the first two: a reasonable expectation for most programs and another complication that *pro forma* rating would have to accommodate.

Significant evidence against the social promotion possibility is found in the analysis of the HERI survey data summarized in Figure 6, which shows a similar pattern to the DASL averages in Figure 4: grades predictive of writing and of writing development, but this time using student perceptions instead of faculty perceptions.

In summary, we cannot entirely reject the idea that some of the rating variance is associated with faculty expectations tied to student seniority or class level; however, the weight of evidence shows that this cannot be the predominant effect. We conclude that the average ratings indicate how well students are meeting general expectations of faculty with respect to writing in disciplines, and that in the aggregate, the ratings represent a broad writing construct.

6.2 Generalization and Extrapolation

Claim 2: The regression models of DASL ratings describe average development of college writers over four years at Furman University.

A naïve description of college is that students spend four years accumulating academic knowledge and skills—and that if we graphed learning over time, the lines would slope up. Translating this picture into a measurement context is difficult. Maybe students are not learning much (Arum & Roksa, 2014), or maybe our model of the process is too simple to detect change (Sullivan & McConnell, 2018) by omitting important inputs. We assume that learning is a product of not just time in class, but personal factors like engagement with the material and prior academic experience, as well as external factors like the discipline being studied.

Astin's (1984, 1996) model of student development is a good fit for the DASL ratings because of its generality: outputs are determined by inputs and environmental factors. We assume that continued engagement with writing practices in college courses leads to better writers over time but that this progress is mediated by environmental and personal factors. The regression models of the DASL ratings specify a relationship between the outcome (writing ratings), the personal inputs (demographics and abilities), and environment (exposure to college over time, course subjects, and raters). The parameter estimates in Tables 5 and 6 quantify the relationships. As such, the model is general enough to apply to any learning outcome that may change over time.

Within the context of Furman University, the regression model fit is a reasonable assessment of the generalizability of the findings. The marginal R^2 of .297 in Table 6 is a measure of model fit only using the fixed effects (the predictors with estimates at the top of the table) and is about half the size of the conditional R^2, so about half of the understanding of ratings comes from trends (time and grades as predictors) and half comes from the individual characteristics of students, raters, and subject, which together comprise a swath of inputs and environmental factors.

A visual comparison of the empirical statistics to the models in Figures 4 and 5 shows a good fit to average values of scores (linear model) and probabilities (binomial models). The model coefficients in Tables 6 and 7 have convincingly non-zero estimates and provide a good average guide to interpreting the modeled averages and probabilities.

In comparison to the R^2 of .297 here, a study of 210 portfolios in Kelly-Riley et al. (2016) modeled a holistic rating using high school GPA and standardized test scores (ACT or SAT) with R^2 values of .06 to .09. Since R^2 depends on how much score variation there is, we would expect that a longitudinal study like the current one would have more score variability and hence a larger R^2.

The relatively low R^2 values and rating reliability, contrasted with good estimates of model parameters, tell us that the averages are predictable, but that individual student developmental paths are not (the shaded standard deviations in Figures 4 and 5 illustrate this). And those averages have interesting properties: they indicate learning over time, but not in the same way for every student type; there are evident disparities. In particular, the Term coefficient in Table 6

indicates a .16 average gain per term for students with an rGPA of zero (average-grade earners), predicting a four-year gain of 1.3. Probability estimates in Table 7H give a 50 percent chance that an average-grade earner receives a 5 rating ("ready to graduate"). The rGPA interaction term adjusts the growth rate by .07 per term per grade point above or below average. It is the interaction term that produces the divergent plots in Figure 4.

Arum and Roksa (2014) analyzed scores from the Collegiate Learning Assessment (CLA), a standardized test of critical thinking, complex reasoning, and writing. Score distributions by college selectivity in their Figure 2.3 (N = 1,666) were used to generate our Figure 7, which shows mean scores of college freshmen and seniors.

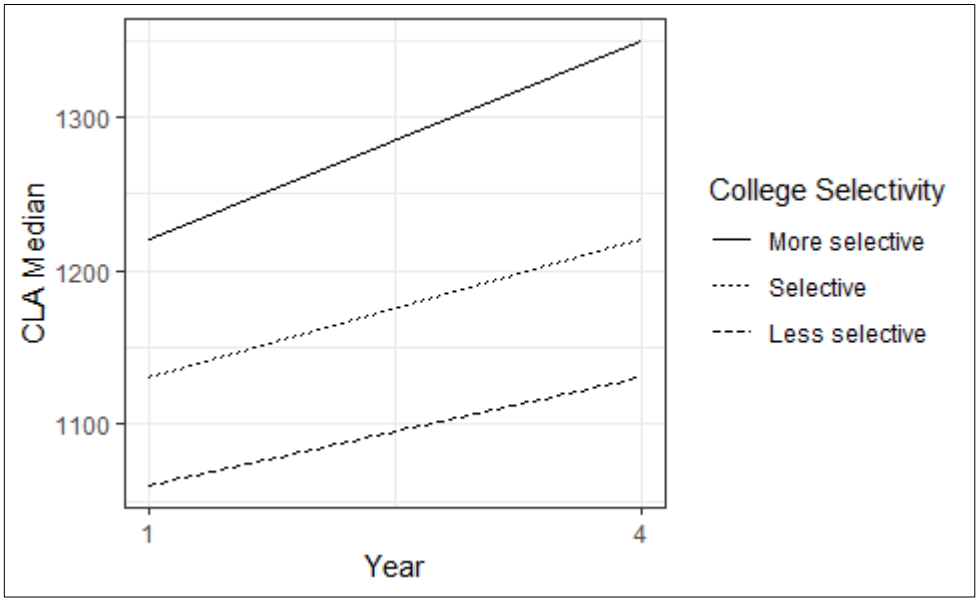

Figure 7

Data Reproduced from Arum & Roksa (2014, fig. 2.3; N *= 1,666), Showing Median CLA Scores by Year in College and Selectivity of the Institution, with Lines as Visual Guides to Each Type*

Note. Values were approximated visually from the original and are provided here as a convenience to the reader. The original gives interquartile ranges, showing substantial overlap in score ranges.

The pattern shows that students at less selective colleges start with lower scores on average and graduate at levels equivalent to where the selective college students start. The same pattern repeats for "more selective" institutions. Selectivity is largely based on demonstrated academic ability prior to admission, analogous to the rGPA explanatory variable in our models, but at a system level instead of institutional level. Arum and Roksa (2014) describe a divergence in measured skill levels: "Students who enter college with higher SAT scores gain more on the CLA" (p. 44).

Within an institution, and certainly within selectivity classes of institutions, there is a wide variation in the academic preparation and performance of students. However, on average, more selective institutions recruit students with higher academic qualifications. Therefore, the trend in Figure 7 is a system-scale version of the divergence seen in the DASL averages found in Figure 4. Therefore, it is reasonable to conclude that the Arum and Roksa (2014) findings support the "growth with disparity" findings here.

Still, both DASL and CLA averages are abstractions of learning that may have different measurement properties at different times. For example, it is possible that the measurement scales have more power to detect differences for college seniors than for college freshmen. In that case, the fan-shaped divergences in Figure 4 and Figure 7 may be illusory and show differences between groups that are actually constant over time. These questions linger because of the lack of a "gold standard" measure.

There is, however, such a measure available as an outcome of college. Salaries of college graduates are not a particularly good proxy for writing ability, but they have the advantages of being eminently measurable and highly consequential. The data come from Chetty et al. (2017), in which they "characterize intergenerational income mobility at each college in the United States using data for over 30 million college students from 1999-2013." Their report received wide coverage, including custom pages on the *New York Times* website for most colleges to illustrate the economic mobility of its graduates.

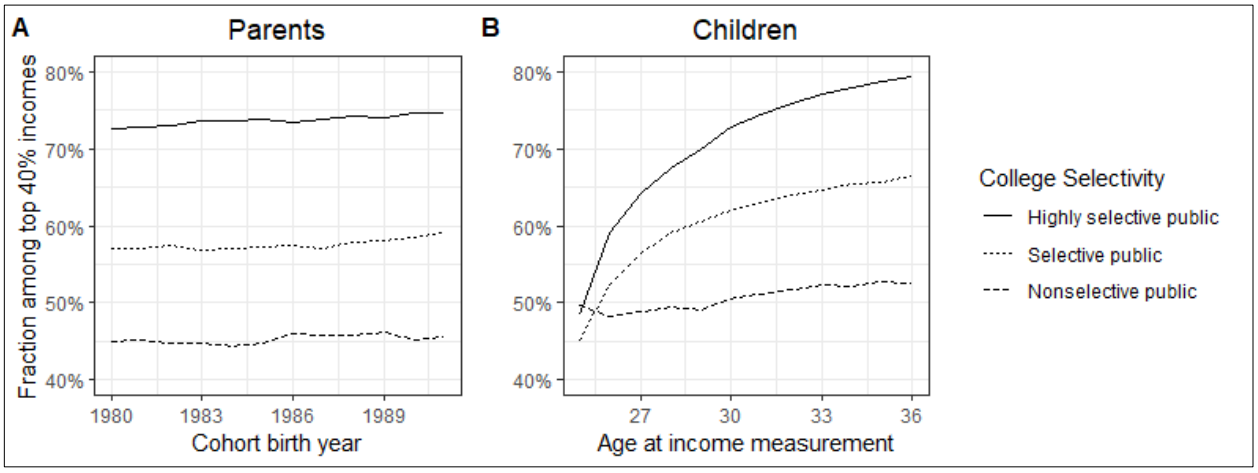

Figure 8

Data from the Equality of Opportunity Project (Millions of Samples), Showing Average Parent and Child College Graduate Incomes in the Top 40 Percent of Earners Nationally, Disaggregated by the Selectivity of College Attended (Publics Only)

Note. *Plot A shows a near-constant relationship between parent incomes and selectivity, and plot B shows each cohort's membership in the top 40 percent of earners of graduates by college selectivity and age when incomes were measured.*

The plots in Figure 8 were calculated from the summary data source included in the supplementary material in Chetty et al. (2017) and are similar to plots in the original report (particularly Figure IIa). The flat lines in Plot A show a constant tendency for wealthier parents (those in the top 40 percent of earners) to send their children to more selective colleges. The diverging lines in Plot B show that the earnings outcomes of their children start with no disparity at graduation but diverge quickly after according to the selectivity of the college. The graphs are ecological (showing separate cohorts who happen to be at different times) and not truly longitudinal. Despite this, interpreting the graphs as an approximate average trend in earning ability in the years after graduation is reasonable. By implication, demonstrated academic ability (via high grades and test scores) in high school, combined with the resources of wealthier parents, leads to a significant divergence in outcomes after graduation. Here, there is no question of scale compression (the salaries are adjusted for inflation).

It is reasonable to assume that the individual traits of students, including the effects of wealth or poverty, combined with the characteristics of the college environment (including those associated with selectivity), have effects on the salaries of graduates. At a national level, and across time, we have seen divergence in self-assessments (Figure 6), standardized tests (Figure 7), and incomes (Figure 8). We assume that these are related, and if this stratification happens at the system level, we conclude that stratification happens *within* universities as well.

Even if we accept that some college students turn out to be better writers than others, should we believe the shape of the graphs in Figure 4, which suggest that the gaps grow over time? Recall that scale compression is a threat to the validity of that conclusion. The Chetty et al. (2017) study and other divergence patterns are suggestive but cannot answer that question definitively. But if writing disparities in college are constant, rather than expanding, then how did the gaps come to be? Were students born with them, or did they expand over time, only to become fixed during college? As a formal exercise, we can divide the gap size in Table 6 (the rGPA coefficient of 0.55) by the rate of increase of that gap (the interaction coefficient of 0.07), and calculate that eight terms before the first college year (i.e., just before entering high school), the writing gap was zero. We make no such claims, but the premise is not absurd: the gaps must originate somewhere, which likely entails continued divergence.

All considered, it seems reasonable to conclude that average DASL writing ratings reveal growth over time, but with disparities in student development. The continued expansion of these disparities in college cannot be proven with the current data set, but the modeled averages are analogous to other expanding gaps systemic to higher education, and it is the simplest explanation to fit the data. We suggest that there is some social promotion that overstates growth, and some scale compression that overstates divergence, but that despite those sources of error, the data contain important clues to writer development in college: there are significant gaps in performance at graduation, these are predicted by grade averages, and—at the least—performance gaps are not shrinking in college; they are more likely to be expanding.

6.3 Implications

Claim 3: Observational ratings are a useful supplement to other types of writing assessment.

The findings here demonstrate that it is feasible to analyze student learning development by means of easily obtained observational ratings by experts. This claim is supported by the previous discussion of reliability and validity, within the limitations described there (for example, the DASL ratings are not shown to be appropriate for judgments about individual students). There are two advantages to observational ratings. One is that much larger sample sizes can be generated than with manual rating methods of individual papers. The second is that since "writing ability" is a social construct, not one pulled from nature like electromagnetic force, a social "crowdsourced" estimation of writing ability may be closer to ordinary meanings of the phrase than the results of standardized instruments are. For most purposes concerning a student's long-term success, it is the social reception of writing that matters, like a boss' approval, article publication, books sold, or the impact of a letter or report.

On the other hand, the DASL ratings are not a tool to cleanly separate writing from the general academic ability. They cannot serve the same purpose as analytic rubric analyses that examine qualities of written works, like style and organization. Because of this limitation, we suggest that the observational ratings are a useful complement to, not replacement of, standardized measures.

Are the DASL ratings fair? Are they biased against vulnerable student populations? The rating method relies on professional experience and direct observation over several weeks. The synthesis that produces ratings is subjective, and subjectivity is sometimes disparaged in measurement because of the inevitability of biases that affect individual ratings. Avoiding such bias is one of the attractions of standardized measures, but of course, those can be biased too: just systematically instead of individually. See Tough (2019) for a compelling account of the SAT's bias against low-income students and Poe et al. (2018) for a broad account of social justice and writing assessment. Since we are not claiming that DASL ratings should be used for decisions about individual students or even used for individual student feedback, the discussion of consequences and fairness here only concerns the systematic use of the DASL ratings to make decisions about groups of students.

Individual rater average biases are already accounted for in the hierarchical model, as are subject-specific effects. The question we consider here is whether subgroups of students are systematically receiving higher or lower ratings that may represent systemic unfairness to the extent that it would affect the findings. To assess that question, we considered gender, race, family income, first-generation status, and athletes. Of these, the regression models found small negative effects for males and non-White students, each amounting to a quarter of an average term's development. A larger effect favoring non-athletes amounts to nearly a full term's development. The most disadvantaged group of students identified is therefore non-White male athletes, who are predicted to be slightly more than a term behind their White female non-athlete peers in average writing development, in addition to any grade-induced differences. There is not enough information in the data set to know whether these imputations reflect real differences in these groups that additional explanatory variables would account for, or if there is systematic

rating bias against these groups. For the present discussion, the question is, If these effects represent unfairness, are they statistically significant enough to affect the conclusions? We estimated the size of the effects by modeling White non-athletes separately from the complementary group (non-White or athlete). The regression coefficients from these tests gave essentially the same values as the aggregate model in Table 6. The divergence finding is not affected by those student types.

There are plausible explanations why student athletes would receive lower scores. For example, practice schedules and travel to games may limit the amount of time available to do research and write papers. Of course, these coefficients derive from a model that already includes grade averages, which is a significant predictor of writing ratings over time. The real question is, Are grade assignments and the factors that lead to them fair? That question is beyond our scope, but Claim 4 adds context.

Claim 4: Feedbacks related to academic performance and family incomes cause writer development divergence before or in college.

The writing ratings for this study come from a single university, but most universities enroll students at varying levels of academic preparation, family income, and demographics. We suggest that the developmental disparities reported by Furman faculty, and the possibility of divergent development, are likely to occur elsewhere. If so, finding the reasons for this is of general importance.

It is not surprising that writing ratings are associated with grades, but the strength of the association and the divergence effects were surprising. How should we conceive of writing in the context of general academic performance? Certainly, the habits and traits that predict grades (intelligence, prior education, work habits, etc.) would be expected to predict writing ability. At a liberal arts college, written assignments contribute significantly to many course grades, and conversely, the quality of the submissions is related to mastery of course content.

Diverging growth rates in learning have become known as a Matthew effect (those who have more, get more). Stanovich (2009), in a theoretical study of how young learners develop reading ability, poses a feedback loop based on core competencies that can lead to ability divergence over time. Stanovich posits such a "reciprocal causation" that can lead to diverging reading ability, e.g., if the ability to associate words, sounds, and meanings leads to more reading and more reading leads to gains in this phonological skill, there is a positive feedback that can lead to increased learning gaps over time:

> [P]oorer readers often find themselves in materials that are too difficult for them. . . . The combination of lack of practice, deficient decoding skills, and difficult materials results in unrewarding early reading experiences that lead to less involvement in reading-related activities. Lack of exposure and practice on the part of the less skilled reader delays the development of automaticity and speed at the word-recognition level. (p. 25)

Explaining the results in the present study may entail a similar feedback mechanism, for example, student attitudes about writing (like fixed versus growth mindset), where a "poor"

attitude leads to poor results, which sustains or worsens the attitude, leading to less effort. On the other hand, a "good" attitude may similarly lead via good results to positive feedback and hence to sustained or improved attitudes. Given the HERI results on self-assessed writing shown in Figure 6, feedback between attitudes and writing performance is not far-fetched.

The connection between grades and learning could also include the role of work habits. The HERI freshman survey asks how often students "failed to complete homework on time," and most Furman students respond with "not at all" or "occasionally." A *t* test of the linked DASL writing ratings taken in the first year of college shows an average rating gap of .15, favoring those who say they always do homework on time ($t(2,889) = 4.82$, $p < .0001$). That gap is still .14 in the senior year ($t(2,743) = 3.86$, $p < .0001$), about one term of development according to Table 6, which has the Term coefficient at .16. These are clues that outcome disparities are amenable to quantitative research that includes student attitudes, behaviors, and beliefs as well as socio-economic histories.

We do not have a fully-developed model to support Claim 4 and must leave it as a strong possibility requiring further research.

7.0 Conclusions

We propose four contributions from this work to the theory and practice of assessing writing. First, the observational method of data-gathering succeeds at generating large samples of useful data. It is a natural extension of portfolio review, with one more step toward trusting teachers and a step away from standardization. We do not advise that DASL-type ratings be used for decisions, advice, or feedback for individual students.

Second, the introduction of explanatory variables in hierarchical ordinal regression, as in Section 5.6 is not new, but the example here may lead to productive work integrating Rasch-type methods into Astin-type models of student achievement. The use of time as an independent variable offers the possibility of invariant developmental scales, and the addition of grades, demographic variables, or other student descriptors expands our ability to understand causes.

Third, the discovery of divergent writing development at our university combined with national data suggesting generality, poses serious problems for how we think of college education. Does college mainly benefit those who earned the best grades in elementary and high school? Are the second- and third-tier grade earners subsidizing the education of high GPA students (via merit aid) while receiving mediocre outcomes?

Finally, the importance of grade averages in predicting average writing scores suggests the need to better link grades and learning assessments. We recommend that writing studies include student identifiers associated with each score and associate those with student GPAs and other explanatory variables like time in college. This was done in some of the writing studies we reviewed, but the use of grades is avoided in most assessment offices because of accreditation proscriptions. This is a mistake; institutional self-improvement efforts should integrate grading with other types of assessment to understand both.

To end on an optimistic note, we observe that there is great variation among students within the same GPA quintile. Grades and socioeconomics are not deterministic, and we can learn from students who succeed despite the odds. Paul Tough's (2019) *The Years that Matter Most* is good preliminary reading.

8.0 Directions for Future Research

The R^2 values in Table 6 tell us that important variables are missing from the model, comprising variance yet to be explained. Student attitudes, beliefs, and behaviors are a good place to start. The national HERI survey data undoubtedly contains more insights. While the data we used are now many years old, newer data exists and can be obtained from HERI by writing a research proposal. Likewise, we still have much to learn from the vast salary data set from Chetty et al. (2017). Models of institutions can combine that data with public data from the National Center for Education Statistics and other sources to expand our system-level understanding of outcomes. For example, how are the disciplines studied by graduates related to economic mobility? See Hill and Pisacreta (2019) for a step in that direction.

Since grade averages at our university remain largely constant over four years, the relationship to DASL ratings suggests a feedback mechanism as described by Stanovich (2009) whereby persistent student characteristics reinforce learning growth. As noted, student attitudes are a candidate for such a feedback mechanism, but there are others. Given the importance of grades and grading in progress toward a degree and as an advertisement of learning on transcripts, it is imperative that colleges and universities take research on the subject seriously. Millet (2010) showed that reporting grade statistics back to instructors increased reliability, one measure of fairness. We suspect that comparative summaries of DASL ratings would similarly improve rating reliability, but that has not been done yet.

Finally, the DASL project should be replicated at other institutions to see how generalizable the findings are. In addition to research, the results can be used for institutional reporting of learning outcomes.

Acknowledgments

We would like to thank Norbert Elliot for his insights and helpful feedback on the literature review. Anonymous reviewers provided insights that greatly improved the manuscript, including a better adherence to Kane's validity model and in prompting us to look for external data: that's when we found the HERI, CLA, and salary comparisons. The DASL data set for this study would not have been possible without the cooperation of the wonderful faculty of Furman University, nearly half of whom contribute voluntarily each term.

Author Biographies

David Eubanks serves as Assistant Vice President for Institutional Effectiveness at Furman University.

Sara Vanovac serves at Associate Director of Institutional Research at Furman University.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Wiley & Sons, Inc.

Anderson, P., Anson, C. M., Gonyea, R. M., & Paine, C. (2015). The contributions of writing to learning and development: Results from a large-scale multi-institutional study. *Research in the Teaching of English*, *50*(2)199-235.

Arum, R., & Roksa, J. (2014). *Aspiring adults adrift: Tentative transitions of college graduates*. University of Chicago Press.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of College Student Development*, *24*, 297-308.

Astin, A. W. (1996). Involvement in learning revisited: Lessons we have learned. *Journal of College Student Development*, *37*, 123-133.

Beatty, A. S., Walmsley, P. T., Sackett, P. R., Kuncel, N. R., & Koch, A. J. (2015). The reliability of college grades. *Educational Measurement: Issues and Practice*, *34*(4), 31-40.

Behizadeh, N., & Engelhard Jr., G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Carbonneau, K. J., Van Orman, D. S., Lemberger-Truelove, M. E., & Atencio, D. J. (2019). Leveraging the power of observations: Locating the sources of error in the Individualized Classroom Assessment Scoring System. *Early Education and Development*, *31*(1), 84-99.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1-32. https://doi.org/10.18637/jss.v076.i01

Casbergue, R. M. (2010). Assessment and instruction in early childhood education: Early literacy as a microcosm of shifting perspectives. *Journal of Education*, *190*(1-2), 13-20.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). *Mobility report cards: The role of colleges in intergenerational mobility* (No. w23618). National Bureau of Economic Research.

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560-575. https://doi.org/10.1111/medu.12678

Engelhard Jr., G., & Wind, S. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments.* Routledge.

Eubanks, D. (2017). (Re)Visualizing rater agreement: Beyond single-parameter measures. *The Journal of Writing Analytics*, *1*, 276-310.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Higher Education Research Institute. (2006). *College senior survey trends database*. https://heri.gseis.ucla.edu/archives/

Hill, C. B., & Pisacreta, E. D. (2019). *The economic benefits and costs of a liberal arts education*. The Andrew W. Mellon Foundation. https://mellon.org/media/filer_public/82/fa/82fac4d2-8e1c-4b7e-ba80-5efbd396c6c9/catharine_hill_on_economic_outcomes_1-9-2019.pdf

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187-200.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527.

Kane, M. (2006). Validation. In R. Brennen (Ed.), *Educational Measurement* (4th ed., pp. 17-64). American Council on Education.

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177-182.

Kelly-Riley, D., Elliot, N., & Rudniy, A. (2016). An empirical framework for ePortfolio assessment. *International Journal of ePortfolio*, *6*(2), 95-116. http://www.theijep.com

Martinez, C. T., Kock, N., & Cass, J. (2011). Pain and pleasure in short essay writing: Factors predicting university students' writing anxiety and writing self-efficacy. *Journal of Adolescent & Adult Literacy*, *54*(5), 351-360.

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education*, *65*, 48-60.

Millet, I. (2010). Improving grading consistency through grade lift reporting. *Practical Assessment Research and Evaluation*, *15*(4), 1-8.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(2), 5-12.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading & Writing Quarterly*, *19*(2), 139-158.

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in *Assessing Writing*. *Assessing Writing*, *42*, 100418.

Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*. WAC Clearinghouse.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *The Journal of Writing Analytics*, *1*, 227-275.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420.

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of Education*, *189*(1-2), 23-55.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743-762.

Sullivan, D. F., & McConnell, K. D. (2018). It's the assignments—A ubiquitous and inexpensive strategy to significantly improve higher-order learning. *Change: The Magazine of Higher Learning*, *50*(5), 16-23.

Tough, P. (2019). *The years that matter most: How college makes or breaks us*. Houghton Mifflin Harcourt.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. University Press of Colorado.