Building a Transcript of the Future

Benjamin P. Koester
University of Michigan
Department of Physics
450 Church St.
Ann Arbor, MI 48109
bkoester@umich.edu

James Fogel
University of Michigan
Department of Economics
611 Church St.
Ann Arbor, MI 48109
jsfog@umich.edu

William Murdock III
Harvard University
Department of Economics
1805 Cambridge Street
Cambridge, MA 02138
wmurdock@g.harvard.edu

Galina Grom
University of Michigan
Department of Physics
450 Church St.
Ann Arbor, MI 48109
grom@umich.edu

Timothy A. McKay
University of Michigan
Department of Physics
450 Church St.
Ann Arbor, MI 48109
tamckay@umich.edu

ABSTRACT

The pathways and learning outcomes of university students are the culmination of numerous experiences inside and outside of the classroom, with faculty and with other students, in both formal and casual settings. These interactions are guided by the general education requirements of the university and by the learning goals of the student. The only official record and representation of each student's education is captured by their academic transcript: typically a list of courses described by name and number, grades recorded on an A-F scale and summarized by GPA, degrees awarded, and honors received. This limited approach reflects the technological affordances of a 20th century industrial age. In recent years, scholars have begun to imagine a transcript of the future, perhaps combining a richer record of the student experience along with a portfolio of authentic products of student work. In this paper, we concentrate on first, and develop analytic methods for improving measures of both classroom performance and intellectual breadth. In each case, this is done by placing elements of individual transcripts in context using information about their peers. We frame the study by addressing basic questions. Were the courses taken by the student difficult on average? Did the individual stand out from their peers? Were the courses representative of a broad intellectual experience, or did the student delve into detail in the chosen field of study? And with whom did they take courses?

CCS Concepts

•**Applied computing** → **Computer-assisted instruction**; •**General and reference** → *General conference proceedings*; •**Networks** → *Network economics*; •**Information**

This work is licensed under a Creative Commons Attribution International 4.0 License.

Copyright is held by the owner/author(s).

LAK '17 March 13-17, 2017, Vancouver, BC, Canada

© 2017 ACM. ISBN 978-1-4503-4870-6/17/03.

DOI: <http://dx.doi.org/10.1145/3027385.3027418>

systems → *Data analytics*;

1. INTRODUCTION

An academic transcript summarizes the career of a student. University registrars record and store the grades, degrees, courses, and other credentials that students seek when they enroll [10]. Students may then choose to share them with institutions where they continue their studies, or with potential employers. As such, the transcript serves first and foremost as an official validation that a student did indeed participate in academic activities at the university (took classes, earned a degree). It also forms a learning profile of the student; a representation of the education they have received. The courses, grades received in those courses, a degree, and some indication of the area of study provide our only record of the way each student met the educational goals of the institution (e.g. [19]). Institutions use their collection of transcripts to better understand the progress of students through their campuses. Individuals charged with evaluating students - typically graduate or professional school admissions teams and potential employers - learn to read the genre of the transcript with care, attempting to glean deeper information about academic performance, intellectual breadth, disciplinary depth, and effort. Often this interpretation of the transcript becomes baroque, creating a narrative from crumbs of information. Students, too, react to the transcript, attending closely to things which we record (like GPA) while discounting those we don't (intellectual breadth or effort).

Clearly the traditional transcript paints an incomplete picture of the student experience both in and out of the classroom ([2]). Hence, many have advocated for an e-portfolio [18, 20] combining information about courses and co-curricular activities [9] with content produced by the student, all with the official stamp that accompanies a transcript. For the student, a portfolio that grows over an academic career can engage the individual in reflection on progression toward future goals. The process of metacognition is known to enhance learning [4].

As we work toward a rich transcript of the future, we be-

gin by asking basic questions. Absent appropriate context, many elements of the traditional transcript are difficult to interpret. For instance, were the courses taken difficult? In what areas, and to what degree, did this individual stand out from their peers? Were the courses taken representative of a relatively broad intellectual experience, and how deeply did this student delve into detail in the chosen field of study? With whom did they take courses - were they isolated with a homogenous group, or did they interact with a diverse cohort? Even today, advances in the availability and processing of information allow each transcript to be placed in context, becoming a richer, more informative portrait of a student. Relatively simple calculations, based entirely on current administrative data, can provide substantially refined measures of relative student performance, along with more insight into the nature of the courses taken and diversity of classmates.

1.1 Use of Grades in the Transcript

Letter grades are the common currency of academic success in higher education [23]. While their apparent similarity suggests that all A's are equal, challenges to their effective exchange emerge at many levels. Differences among disciplines in content, evaluative style, grading practice, and strength of students make grades awarded in different classes difficult to compare [12, 14]. Reward systems in higher education use grades as incentives in ways which further distort their utility for both students and faculty [1] and periodic attempts to regularize grades across courses and disciplines have found little success (e.g. [22]). Some authors have been sufficiently troubled to declare the use of grades unethical: "inconsistent grading policies render the competitions for academic awards unfair, deprive people of positions they merit, and leave people, students, institutions, and societies less well off than they would otherwise be." [14].

Insofar as grades continue to be an integral part of the transcript, derivatives of them will continue to be used as summary statistics. The grade point average (GPA), typically a credit-hour weighted-average, is the most widespread. On its face, GPA cannot be a measure of how much a student has learned: it does not grow when a student takes more courses. Indeed, it may fall as a student learns more. Imagine the dismay of the 4.0 GPA student receiving their first B+ in a challenging junior level class. Despite learning more - perhaps a lot more - their GPA has fallen [16]. Instead, GPA serves the same dual purpose present in all grades. It is considered both as a summary of a student's mastery of course material and a tool for ranking and locating each student among their peers. Both purposes are undermined by inconsistency of practice: mean grades given out in courses at the University of Michigan vary by more than 25% across the disciplines. These variations provide false signals, suggesting to all students particular talent in courses with high average grades and a lack of ability in courses with low average grades. Ranking too is harmed. Just as in Meyer's time [17], a student may obtain the same high GPA through exceptional performance in courses giving low average grades or average performance in classes giving high average grades. Perhaps the simplest paliative for grading dispersion is to report a mean course grade alongside each student's grade on a transcript. This approach was adopted, for example, at Dartmouth in 1994 [6], but it does nothing to alter GPA or

affect its primary uses.

In the outside world, GPA is also used in several ways. In some contexts, it is seen as a measure of tenacity. Those who achieve a high GPA have demonstrated a sustained ability to meet whatever requirements have been placed on them. In other contexts, GPA is used as a predictor of future performance. Those with high GPAs are expected to perform well in subsequent endeavors. Is GPA adequate for these purposes? Perhaps. Employers, graduate and professional admissions committees, and those administering University honors of various kinds have used GPA for a century, and found value throughout [13]. But there are real reasons for concern, and several alternatives to GPA have been considered, most of which acknowledge the considerable effort expended to assign grades, but suggest that the information is best applied in a relative sense [14].

1.2 Alternatives to GPA

Practical alternatives to the GPA begin with grades as reported by instructors, then utilize information contained in the distribution of grades to refine measures of ranking [25, 15, 5, 11, 3]. Larkey and colleagues [15] were among the first to use basic linear models in an effort to extract measures of student achievement. Caulkins et al. [5] presented several methods that recompute GPA after normalizing student grade distribution according to other students in the same course. In essence, each of these takes existing grading data as-is and calculates a metric that is less sensitive to the peculiarities of grading practices among courses and more indicative of student performance. Perhaps the most sophisticated of these efforts, a Bayesian latent trait formulation of a summary statistic to replace GPA-based measures [12], was considered for use by the faculty at Duke, but ultimately rejected [8]. So far, none of these schemes has entered widespread use. In what follows we consider several possible replacements for GPA.

1.3 Courses and Classmates

The transcript also provides insight into the breadth and depth of each student's course of study. Readers of a transcript place the list of courses a student takes into context using a combination of their experience and imagination. Institutions might provide much more precise insight, clarifying how students met their graduation requirements. Institutions might also place the intellectual breadth and disciplinary depth of courses taken by a student in local context, comparing each student's transcript to those of their peers, both within their discipline and across the institution. In an important sense, the list of courses taken by a student provides a measure of their success or failure to meet the goals of a liberal education, regardless of performance. It also encodes information about the student's creativity in meeting course distribution requirements, whether that means exposure to many new topics or detailed study in a few.

Institutions which have access to the transcripts of all their students can use this information to add important context to what they present for each student. They can refine their measures of student performance, perhaps augmenting or replacing GPA. They can also place the course-taking and enrollment patterns in context, along with the diversity of individuals with which the student took classes.

A student taking courses that expose them to peers from many different majors, backgrounds, and identities likely has a much different experience from one who primarily interacts with students like them in course of study, background, or identity.

Advances in the use of information technology in education have revitalized the use of data in evidence-based advising and decision making [3]. A new field of Learning Analytics has emerged, working to use data about learners and their environments to understand and optimize learning [24]. Perhaps the time has come to seriously reconsider the ways in which we represent a student's performance in college. In this work, we explore a few small steps toward the transcript of the future. Keeping in mind the intent of the transcript - to accurately summarize and represent a student's education - we explore several alternatives to the GPA and their sensitivities, as well measures of intellectual breadth and diversity of experience.

2. DATA

The University of Michigan Data Warehouse contains grades, student admissions information, and demography back to Fall 1974. In this work, we are chiefly concerned with grades, which are complete back to at least Fall 1998. Thus, we consider student course data from Fall 1998 through Winter 2014 on the Ann Arbor Campus. This data set includes 4,384,169 grades for 192,987 students, who received grades in 15,571 courses distributed among 318 subjects. We only consider grades for students that took a course for credit and were still enrolled at the end of term. We ignore withdrawals, incompletes, audits, and pass/fail evaluated courses. For students that took a course multiple times, we consider only the last attempt, mirroring the U-M's standard practice used to compute a GPA. This gives a grand total of 132,223 course-terms over 16 years.

Where we are concerned only with student grades, our student sample is quite liberal. Students who transfer into and out of the University and across its many colleges are retained. Students who arrived on campus before Fall 1998, or who had not yet completed their degrees in Winter 2014 are also included. Some statistics require a student's complete undergraduate record for calculation. These include measures of academic performance at graduation and the diversity of subjects and classmates with whom the student interacted. For these statistics, we further limit our data set to include only students admitted as freshmen since Fall 2005, who took at least 30 courses on campus, and graduated.

At Michigan, majors of students are only finalized upon graduation; they are not expected to declare an intended major before matriculation to campus, and usually declare an intent only later in their sophomore year. As at most universities, instructors award letter grades which are then converted to grade points: an A = 4.0, A- = 3.7, B+ = 3.3, ..., E = 0. We note that the grade points given for 'plus' grades to Business School undergraduates are unusual. A+ = 4.4, B+ = 3.4, and so forth. In this analysis, these are normalized to the grade points given by the rest of the University. We also only consider undergraduate courses at the University of Michigan, which are numbered between 100 and 499. 100

and 200 level courses are mostly introductory lectures, discussion, laboratories, and seminars, and they contain among their ranks the highest enrollment courses on campus. By contrast, 300 and 400 level courses are typically directed at majors and have smaller enrollments formatted as small lectures, laboratories, and research for credit. Course types and size vary dramatically, ranging in enrollment from 1 to more than 2000 in a term and in structure from apprenticeship to massive lecture. Course content spans the intellectual range of the university, and both instructional style and grading practice vary with discipline.

2.1 Placing Student Performance in Context

A typical transcript reports a single student's list of courses taken and grades received without context. Little information is provided about what the courses were like, who took and taught them, how they were taught, evaluated, or graded. Lacking this context, the transcript is difficult to interpret. Given that registrars possess information adequate to provide substantial context, we argue that it should be put to use.

Consider the list of the courses a student took and the grades they received. Since courses are listed only by name and number, the reader must guess at content studied, work done, nature of the instructor or classmates, evaluative style, or grading practice. If we consider instead the context available from the full collection of student transcripts, substantially more information is available. It becomes possible to compare this student's performance to others in this class, even to relevant subsets of others (e.g. those who continued on to complete the same major). We can learn more about who takes this class; what they take before or concurrent with it, what they continue on to study, and what degrees they ultimately receive. We can see whether students in this class typically receive grades higher or lower than they get in other courses. Analysis of the full network of courses and student grades allows us to knit together performance comparisons even between students who were never coenrolled in a class. In this context, grades and courses taken may be viewed through a different lens. Of course the quality of these relative performance measures depends on how deeply each student is embedded in the campus network of coenrollment. As a result, we begin with some descriptive exploration of how many classmates each student in our sample has encountered directly.

Undergraduate degree-seeking students need 120 credits to graduate, which typically requires taking between 30 and 50 courses. Usually, before declaring a major, they take an array of introductory and prerequisite courses which include classes both small (writing courses and seminars) and large (lecture classes with enrollments from 50 to 2000). Most direct comparisons among coenrolled students take place in these large lecture courses. To give some sense for the numbers of classmates encountered by each student during their career, Figure 1 shows the number of total classmates for sets of students that, at any point in their career, took one of a selection of introductory courses at Michigan. These courses are part of the Psychology, English, Chemistry, and Statistics Majors, but are taken by large numbers of non-majors as well. Over their careers, the students in our sample accumulate anywhere from a dozen or so classmates to nearly

25,000, with typical values ranging from 7,500 to 12,500. Grades handed out in these head-to-head comparisons form the basis of the grading system and the grade point averages that characterize students’ performance over their career.

Figure 1: Histogram of the total number of unique classmates a student had during their academic career, given that they took one of four selected large courses at the University of Michigan.

Patterns of course-taking also create clusters of students who take a series of large introductory courses together, allowing us to repeatedly compare the performance of these classmates at large nodes in the student network. Figure 2 shows the course correlation matrix for the 40 University of Michigan courses with the largest average enrollments. It is no surprise that courses cluster within subject (e.g. CHEM, PHYSICS), or that there is a general clustering of science courses. There are also courses (e.g. PSYCH 111) with little association to core science courses, and in general no particularly strong correlation with any other high enrollment courses. This is in part due to the diversity of careers of our students, but also to a fundamentally different distribution of course enrollments: science majors take several core high-enrollment classes together with other science majors. Humanities majors experience much more flexible requirements, and hence are less strongly coupled to either science students or even one another, especially outside their major. As a result, more caution is warranted in the comparison of weakly-coupled students than for those who are richly connected.

Finally, the assignment of grades varies broadly among academic departments of the University. Figure 3 answers the question “what kinds of grades do students receive in the course where most grades come from?” To construct this figure, we first identify all departments which teach at least 10 courses with average enrollments of at least 50 students. For those departments and their biggest courses, we

Figure 2: The distribution of course taking. For the top 40 courses by total enrollment, the Pearson correlation coefficients between taking pairs of courses at the University of Michigan. Lighter shades symbolize higher correlation, and hierarchical “complete linkage clustering” dendrograms arrange courses into groups that indicate similar patterns of course-taking.

record the total enrollment and mean grade, aligning the courses from highest to lowest average enrollment and indicating grade in a color scale ranging from red to white as mean grades vary from 2.65 to 3.85 on a four point scale. Grade trends depend strongly on department, level, and division. This reconfirms earlier work of [7], which concludes that the use of grades and grading depends strongly on academic discipline. Interestingly, this was less true in the time of [17], when diversity in the application of new grading standards apparently had more to do with individual faculty than disciplines. Today at Michigan, eight of the ten lowest grading departments are in science, engineering, and math, along with Economics and the Romance Languages. The highest grading departments are more mixed, with humanities departments making up five of the top ten highest, joined by others like Biomedical Engineering and the School of Education.

2.2 Performance Measures Enriched by Context

The general system of grades described above has long been in place and is not likely to disappear soon. Despite its drawbacks, considerable effort is expended in the determination of grades, so that for most courses grades contain useful information both about characteristics of the course and the relative performance of the students who are enrolled. Given signs that grading standards vary significantly across courses and departments, we are concerned that students are punished for taking “difficult”, low graded courses or incentivized to pursue “easy” courses which tend to assign higher grades. Our challenge then is to leverage existing grades to better understand both grading practices in courses and student performance. To this end, we consider four different metrics of average student performance: traditional GPA, two previously suggested standardizations of the GPA, and a fixed effects model which we introduce for the first time here.

Figure 3: Grades at the University of Michigan. For Departments offering at least 10 courses with typical enrollments of at least 50 students, the course catalog number and mean grade (red to white = 2.65-3.85 grade points) are shown. From left to right, the top 10 courses are ranked by enrollment, while from top to bottom, Departments are ordered by their enrollment-weighted mean grades, and color-coded according to division: red for Natural Science and Engineering, blue for Social Sciences, and green for Arts and Humanities

- **Grade Point Average (GPA)** is the traditional weighted-average of course grades, where the weights are credit-hours. Courses are otherwise assumed to be the same and no information about course grade distributions is included.
- **Grade Points Above Replacement 1 (GPAR1)** is a modification of the prescription given in [5]: student grades are compared to the course mean. The original formulation given by Caulkins is modified here - GPAR1 represents a student's career performance and, like GPA, is a credit hour weighted average.

- **Grade Points Above Replacement 2 (GPAR2)** like GPAR1, corrects student grades to the course mean, but then standardizes this difference with the standard deviation of the course. Again, we modify the Caulkins formulation such that a student's final GPAR2 is a credit-hour weighted average of GPAR2s in all courses.
- **Student Fixed-Effects (SFE)** models every grade given to a student as a linear combination of student and course fixed effects, estimated across the full array of student-course-term records. Student i 's grade in course c , term t is written as the sum of a course and term-invariant student component $StudentFE_i$, a student-invariant course-term component $ClassFE_{ct}$, and an idiosyncratic error term, ϵ_{ict} .

$$Grade_{ict} = StudentFE_i + ClassFE_{ct} + \epsilon_{ict}$$

The coefficients in this model may be estimated with by ordinary least-squares techniques.

Our use of the two Grade Points Above Replacement (GPAR) statistics is loosely inspired by Major League Baseball's Wins Above Replacement statistics, which compute "How many more games did your team win with you as a player than they would have with a plausible replacement player?" Here, this question is reformed: "How much higher were your grades than those which would have been received by a plausible replacement student?" In this case, the plausible replacement is the average student in every class you took. Unlike GPA, the two GPAR statistics leverage basic information about the grade distribution in each class. Both take into account the mean grade awarded in the class, ascribing this to the instructor rather than the students, and referencing every student's grade to this local average. GPAR2 also takes the dispersion in grades into account. The intent here is further account for varying dispersion in grades, but this method may backfire, ascribing extraordinary discriminatory power to a course in which almost everyone receives the same grade. Both GPAR statistics, like GPA, fail to account for the possibility that students in some courses may be, on average, much more successful students than those in other courses. Our final model *jointly* estimates student and course effects as parameters in a matrix that couples students to one another through courses taken together. The matrix still contains the full information about the grade distribution in every class, and a solution that returns coefficients encoding the student and course effects should in principle be a more accurate measure of student ability in college courses, at least if the assumptions underlying the fixed effects model are valid. This method does account for the possibility that students taking one course are on average substantially stronger than those taking another.

2.3 Intellectual Depth and Breadth

Collective analysis of course taking patterns for all students can support a variety of measures of intellectual diversity. The distribution of individuals among a set of groups may be described by a class of diversity indices, that take the general form (e.g. [21]):

$${}^qD = \frac{1}{q-1 \sqrt{\sum_1^R p_i p_i^{q-1}}} = \left(\sum_1^R p_i^q \right)^{\frac{1}{q-1}} \quad (1)$$

This is the inverse of the weighted generalized mean. R is the number of unique groups, and p_i is the proportion of members in group i . Setting $q = 2$ makes this inverse of the weighted arithmetic mean, and $q = 1$ is the weighted geometric mean; it reduces to the exponential of the Shannon entropy. When we set $q = 0$, this is the harmonic mean, which just reduces to the total number of groups, R . As q increases, the weight given to the most abundant groups increases.

The variety of subjects of the courses found on a transcript are a proxy for the intellectual breadth and depth of the content that a student was exposed to over his or her career. Using the same sample built for the GPA statistics, we compute a subject diversity index with $q = 2$ for each individual. Here, each subject is considered a group. In much the same way as GPA summarizes grades, the statistic 2D reduces a complex aspect of the transcript to a single number. As with the simple GPA, this statistic makes no explicit account of the student-student network in which an individual is embedded. While a student may have taken a broad range of courses, he or she could have potentially taken them all with students from the same major. Indeed the opportunity to interact with individuals from different intellectual backgrounds is a stated goal of many institutions. This compels the construction of measures that capture the diversity of classmates, that is, people with whom the student had the opportunity to interact. We use graduating major to classify each of a student's classmates, which we call major diversity, wherein we set $q = 2$. This final statistic involves the computation of a large student-student course coenrollment network (e.g. Figure 1), which is the subject of forthcoming paper. To make computation more manageable, we restrict this network and consider only the course coenrollments of students that entered in the College Literature, Science, and Arts (LSA) since Fall 2005 with any other student at the University.

3. RESULTS

3.1 Comparing Single Metric Performance Measures: GPA, GPAR1/2, SFE

The heterogeneity in grades is evident. These grades depend on an interplay of the term the course was taken, the subject, the strength of the students in the class, and the instructor, among other things. GPA is a ubiquitous statistics of a career, agnostic to everything about courses and students aside from the grades assigned. Given that we know mean grades vary by 25% among departments, this cannot be a perfect estimator of student success. GPAR1 attempts to correct for variations in average course grades by comparing student grades to the course means; it is a first-order correction to GPA. Courses with unusually high or low mean grades should skew this grade-based metric less. GPAR2 offers a second-order correction to this effect: deviation from the mean is measured in units of standard deviation, under the assumption that large deviations where the grade distribution was narrow implicitly contain more information than those where the distribution of grade was broad.

In none of these first three metrics is there an explicit account of the strength of the other students in the course. In simpler terms, a grade received in a course with above

or below average students should be interpreted differently than one with average students. This information is put to use in the fixed-effects model. In Figure 4 $\simeq 200,000$ final GPAs and SFEs are plotted. Histograms show the distribution of students in each dimension and individual points show means for courses offered in some departments. SFE, GPAR1, and GPA are all in units of grade points. Increasing GPA generally tracks with increasing SFE but with considerable scatter. That the centroid of the distribution in SFE does not fall on 0 reflects the strong negative skew of the SFE distribution; the departmental mean SFEs more closely match the centroid of the SFE distribution, as they cluster around zero. The GPA distribution is strongly peaked, due in part to the truncation of the GPA scale at 4.0; SFE experiences no such truncation and is spread more broadly, which reflects this measure's better ability to distinguish individual students from one another. A student in the centroid has a GPA of $\simeq 3.5$ and SFE of $\simeq 0.2$. At nearly fixed GPA, departments range from those with higher SFE (Math) to those with lower SFE (English, Psychology). At fixed SFE, GPAs in Math are considerably lower than those in Organizational Studies. While the structure of the two-dimensional distribution is smooth, its composition is rich.

Figures 5 and 6 show similar comparisons. The GPA and GPAR1 track each other well, with considerably less scatter at the centroid than SFE vs. GPA; GPA1 and GPA are more similar in construction than SFE, and hopefully this implies that SFE carries more information. GPAR1 vs. SFE has a scatter somewhere in between the first two figures. In this comparison, the two metrics track each other, and naively one expects symmetry about the line of equality in the distributions. Instead a tilt is apparent: in the lower left quadrant, $\text{GPAR1} > \text{SFE}$, and in the upper right quadrant, $\text{GPAR1} < \text{SFE}$.

Figure 4: GPA vs. SFE for $\sim 200,000$ undergraduate majors at the University of Michigan. Histograms represent total numbers of students in each GPA or SFE range. Points show the means of these quantities for different departments, with a selected few departments labeled.

3.2 Student and Course Effects

The sorting of students by some measure of achievement into different subjects has been commented on many times

Figure 5: GPAAR vs. GPA for $\sim 200,000$ undergraduate majors at the University of Michigan. Contours represent the density of students in each range, and labeled dots show the means of these quantities for various majors.

Figure 6: GPAAR vs. SFE for $\sim 200,000$ undergraduate majors at the University of Michigan. Contours represent the density of students in each range, and labeled dots show the means of these quantities for various majors.

over the years. The situation with the SFE is no different (Figure 7) in its assessment of the typical students in different subjects. For the top courses in total enrollment, the lowest course effect courses have students with higher average SFEs. These top courses are a mix of science (23), humanities (16), social science (13).

Introductory science classes cluster in the low mean course fixed-effect, high SFE quadrant of the plot: CHEM 210,215 (Organic Chem I & II); BIOLOGY 171 (Organismal and Population Biology); PHYSICS 140,240 (Mechanics, Electricity and Magnetism); MATH 115,116,215,216 (Calc I-IV) as well as ECON 101,102,401 (Micro, Macro, Intermediate MacroEcon) and Accounting 271. Intro science labs (BIOLOGY 173, CHEM 211, CHEM 216, PHYS 141, PHYS 241) also contain these high SFE students, but have higher mean course effects as already noted in [1]. In the low SFE, high

Figure 7: The correlation between student effect and course effect for high enrollment courses at the University of Michigan. For each course, the student and course effects averaged over all terms are plotted for Natural Sciences (red), Humanities (black), and Social Sciences (blue).

course effect quadrant are ENGLISH 125, 223, 225, WOMENSTD 220. In general, the course-effects are not unexpected given the grading patterns in Figure 3. However, we caution against interpretation of the SFE as an independent, intrinsic quality of a student, or the course effect as a measure of difficulty and rigor of a course. Student-course interaction is one source of confusion in this picture. Comparing CHEM 210 and CHEM 211, a naive expectation is that the students taking this lecture/lab combination are identical. They have very similar SFE, but receive quite different grades.

3.3 Subject and Major Indices

Figure 8 shows the subject diversity (the diversity of the subjects one studied) and majors diversity (the diversity of majors of one's classmates) indices for all LSA students. The mean of the index, its standard error, and the number of students is given.

In Table 1 the indices are listed for a select set of majors. Several majors – especially Physics and Chem – sit below the mean subject diversity, while Business Administration stands out as particularly diverse; the sheer number of these students with high subject diversity pushes the overall distribution higher. Physics BS and Chem BSChem stand out as high in major diversity, while Psych BA and Business Administration sit on the low end of the distribution.

4. DISCUSSION

Our purpose in this paper is to consider ways to use the full collection of student transcripts to add context to each. We have suggested new ways to estimate performance and an initial method for measuring the intellectual diversity of a student's experience. We conclude by considering a few practical questions which would emerge. We examine the re-ranking of students within and among departments that would occur as a consequence of alternative performance metrics, and interpret the observed variation in intellectual

Figure 8: Subject and Major Indices for Graduating LSA majors. The mean and its standard error are given for each distribution, as well as the number of students.

diversity that emerges college-wide.

4.1 Intra-Department Performance

For three departments in LSA (Figure 9) we consider graduates of those departments in Winter 2012. Each appears with an anonymous ID in both the left column (final cumulative GPA, left axis) and right column (SFE, right axis). A single student is connected between the columns by a line. Crossing of lines indicates reranking within the department. Low Spearman rank correlations (bottom) indicates a greater amount of re-ranking. Within a department drastic re-ranking rarely occurs, with most of the shuffling happening among close neighbors. At least one exception to this trend is student 248591 in Philosophy (middle panel), who fell from the top third of the class in GPA to nearly dead last in SFE. The transcript reveals that this individual earned only about half of their credits at Michigan, all in two years. Most of the courses were 300 and 400 level with relatively high mean grades ($3.1 \geq \langle g \rangle \leq 3.7$). This student

Table 1: Subject and Major Indices by Major. Diversity indices are tabulated for University of Michigan for selected majors that comprise extremes of the distributions.

MAJOR	$\langle D_{sub} \rangle$	$\langle D_{major} \rangle$	N
Psychology BA	4.7+/-0.028	16.4+/-0.088	3169
Psychology BS	5.74+/-0.087	19.8+/-0.27	352
English BA	4.21+/-0.034	17.4+/-0.14	1979
Economics BS	5.68+/-0.083	19.2+/-0.25	452
Bus. Admin. BBA	12.6+/-0.066	13.9+/-0.091	3370
Physics BS	3.63+/-0.096	23.8+/-0.68	124
Mathematics BS	5.42+/-0.069	20.9+/-0.26	663
Chemistry BSChem	3.5+/-0.074	25.3+/-0.49	161

earned a relatively high GPA by receiving below average grades in courses which awarded high mean grades.

4.2 GPA Error

Gross re-ranking within a department is in part a consequence of unusual transcripts, and we hypothesize that local reranking is mostly ‘noise’. Ultimately, GPA and similar measures have an error associated with them that is part systematic and part statistical. The systematic component includes things like a student’s major, which courses a student took, and when they were taken, while the statistical component is driven by how well-sampled is the student’s career; the latter should approach zero as the number of courses taken goes to infinity. Bootstrap resampling provides a simple insight into the magnitude of statistical uncertainty in GPA. For each student, we use bootstrap resampling of courses ($N = 100$) to compute a bootstrap mean GPA and error. The median standard error on the cumulative GPA for graduates is 0.058. Higher GPAs necessitate lower standard errors. This means that in practice, GPAs which differ by less than 0.05 grade points are statistically indistinguishable. This reality is never acknowledged by our system of awards, which attends carefully to the meaningless third decimal place in GPA.

4.3 College Honors

GPA forms the basis for traditional University or College Honors. Students are ranked by GPA and selected accordingly. One concern with this system is that if students accumulate most of their GPA in departments that assign high grades, honors will be biased. Figure 3 suggests that in the highest enrollment classes, which often reach 400-level, there are grading trends among and within departments.

In U-M’s College of Literature, Science, and the Arts, academic honors (called “distinction”) is awarded on the basis of GPA. If this ranking were done by other means, award of these honors would go to different students. Figure 9 hints at the ways distinction might change if we ranked by SFE instead of GPA. The color of the lines indicates how students in departments are re-ranked when LSA graduates are ranked by SFE instead of GPA: students in Physics generally receive higher rankings, those in History lower, and those in Philosophy a mix. The alternative ranking scheme does indeed reshuffle the assignment of distinction. At one extreme of the reordering is English, in which the number of normal, high, and highest distinction students goes from 30, 9, and 9 to 20, 8, and 0 students. At the other end is Mathematics which goes from 13, 10, and 4 to 22, 11, and 15.

Figure 9: The Re-ranking of Students by Department and College: Physics, Philosophy, and History Winter 2012. The final GPA (left) and SFE (right) for students that graduated with an undergraduate degree in Winter 2012 is represented. Anonymized ID numbers match particular students to the text. Lines connect a student between the two columns, and aid in tracking re-ranking within a department. Blue lines indicate that a student was ranked higher in the College of LSA using SFE as a ranking criteria than GPA. Red lines indicate that they fared worse under an SFE ranking. The Spearman rank correlation for intra-department GPA and SFE ranks is given at the bottom as well: lower rank correlations indicate a greater degree of intra-departmental reranking.

4.4 Subject and Major Diversity

As with grading, degree pathways and requirements vary between departments in LSA, and certainly in the University beyond. Graduation requirements exist in part to encourage intellectual diversity in study, and this places a constraint on how much diversity (or lack thereof) is present in a transcript. For instance, Business Administration BBA students have high subject diversity, but upon closer inspection, it turns out that this is one of the few undergraduate programs for which multiple subjects exist within a school: subjects of FIN (finance), STRATEGY, MKT (marketing), ACC(accounting) and several others are all exclusive to the business school. This is in contrast to, for instance, mathematics where all courses are designated MATH. For this reason, it may be that a better measure of subject diversity only depends on the Department that 'owns' a course. Until then, the subject indices may be best considered only within departments, not across.

4.4.1 Major Trends

Individually, the lowest subject diversity indices come from students graduating with degrees in Dance or Art and Design, likely after a transfer from LSA. In fact this describes the bottom 10 in the list. The lowest subject index was 1.40, for a student that took over 50 courses in DANCE, the remaining 11 coming from ENGLISH, WOMENSTD, and an array of singles in other subjects. Interestingly, this student's major diversity index was 23.3, or about 4 points above the mean, which indicates that a broad array

of students from other majors were classmates with this individual. Business Administration students were the highest in subject diversity, with one individual at 24.4. However, 20 of this student's courses come from 13 subjects that are owned by the School of Business. For these situations, the subject diversity becomes a measure intra-department breadth. This same student's major diversity index = 18.46, which is below the mean University-wide.

4.4.2 Outlier Careers

Within a department, these measures are more standardized, and comparable. Overall, subject and major diversity show a mild ($\rho = -0.14$) anti-correlation with cumulative GPA, and have no correlation with SFE (not shown) - our diversity indices provide new, nearly orthogonal information. What is it telling us? As an example (for its large numbers of students) we consider Psych BA graduates. For the 10 highest SFE students, the subject indices range from 3.4 - 7.4. The two students at these boundaries had major indices of 19.1 and 19.2. The student with lower subject index took 37 courses from a total of 11 subjects: 18 PSYCH, 15 SW (Social Work), 7 WOMENSTD and an array of others. The higher index is comprised of 31 courses in 19 different subjects: 10 in PSYCH, and the others spread across the academic spectrum. Neither double-majored. At the other end of the Psych BA SFE spectrum (SFE < -0.75), students range from 4.8 - 8.8 on the subject diversity indices, and 16.7-24.8 on the major indices. The individual at 4.8 took 33 courses in 15 subjects: 14 in PSYCH, 3 in FRENCH and the others spread across other subjects with 1 or 2 instances.

This is in contrast to the 19 subjects taken by the 8.8 student across 37 courses: 10 in PSYCH, 4 in ASIANLANG, 4 POLSCI, and 3 in COMM. These two students were, respectively, 19.1 and 21.1 in the major diversity index.

4.5 Conclusions

This paper explores ways in which existing student transcript information might be placed in context using straightforward techniques. Rather than advocating for any of these measures in detail, we prefer to promote the *idea* of enriched transcripts, and to encourage the community to consider other ways in which we might use existing information to better represent the experience of students on college campuses.

5. ACKNOWLEDGMENTS

This work has been supported by the NSF WIDER grant DUE-1347697 for the REBUILD project, by NSF TUES grant DUE-1245127, and by the University of Michigan Provost's Learning Analytics Task Force through the Learning Analytics Fellows Program. We thank Kar Epker for preliminary work which inspired the diversity analysis. We also thank the U-M Registrar Paul Robinson, the Office of the Registrar, U-M CIO Laura Patterson, and all the staff at the U-M Information Technology Services division for both maintaining and supporting access to this remarkable data set. Finally, we acknowledge the important contributions of former U-M Provost Phil Hanlon and current U-M Provost Martha Pollack. Their strong advocacy of appropriate research using student record data has made learning analytics at Michigan possible. This research has been determined exempt from human subjects control under exemption #1 of the 45 CFR 46.101.(b) by the U-M Institutional Research Board (HUM00079609).

6. REFERENCES

- [1] A. C. Achen and P. N. Courant. What are grades made of? *The journal of economic perspectives: a journal of the American Economic Association*, 23(3):77, 2009.
- [2] A. W. Astin. *What matters in college: Four critical years revisited*. Jossey-Bass, 1993.
- [3] M. A. Bailey, J. S. Rosenthal, and A. H. Yoon. Grades and incentives: assessing competing grade point average measures and postgraduate outcomes. *Studies in Higher Education*, 41(9):1548–1562, 2016.
- [4] J. D. e. Bransford, A. L. e. Brown, and R. R. e. Cocking. *How people learn: Brain, mind, experience, and school*. National Academy Press, 1999.
- [5] J. Caulkins, P. Larkey, and J. Wei. Adjusting gpa to reflect course difficulty, 1996. Working Paper, Carnegie Mellon University, The Heinz School of Public Policy and Management, retrieved on Jan 17, 2017 from <http://repository.cmu.edu/heinzworks/42/>.
- [6] Dartmouth Office of the Registrar. Median Grades for Undergraduate Courses, 2015. Retrieved on Jan 19, 2017 from <http://www.dartmouth.edu/reg/transcript/medians/>.
- [7] R. D. Goldman, D. E. Schmidt, B. N. Hewitt, and R. Fisher. Grading practices in different major fields. *American Educational Research Journal*, 11(4):343–357, 1974.
- [8] B. Gose. Duke rejects plan to alter calculation of grade-point averages. *The Chronicle of Higher Education*, March 1997.
- [9] J. Gutowski. Co-curricular transcripts: Documenting holistic higher education. *The Bulletin*, 34(5), 2006. Retrieved January 19, 2017 from <http://www.acui.org/publications/bulletin/article.aspx?issue=306id=1900>.
- [10] J. Hope. Support campuswide educational goals with transcript enhancements. *The Successful Registrar*, 16(7):1–5, 1 Sept. 2016.
- [11] V. E. Johnson. An alternative to traditional gpa for evaluating student performance. *Statistical Science*, pages 251–269, 1997.
- [12] V. E. Johnson. *Grade inflation: A crisis in college education*. Springer Science & Business Media, 2006.
- [13] E. R. Julian. Validity of the medical college admission test for predicting medical school performance. *Academic Medicine*, 80(10):910–917, 2005.
- [14] C. Knapp. Assessing grading. *Public Affairs Quarterly*, 21(3):275–294, 2007.
- [15] P. D. Larkey and J. P. Caulkins. *Incentives to fail*. H. John Heinz III School of Public Policy and Management, Department of Social and Decision Sciences, Carnegie Mellon University, 1992.
- [16] J. Lorkowski, O. Kosheleva, and V. Kreinovich. How to modify grade point average (gpa) to make it more adequate. In *International Mathematical Forum*, volume 9, pages 1363–1367, 2014.
- [17] M. Meyer. The grading of students. *Science*, 28(712):243–250, 1908.
- [18] R. Miller and W. Morgaine. The benefits of e-portfolios for students and faculty in their own words. *Peer Review*, 11(1):8, 2009.
- [19] A. of American Colleges and Universities. The leap vision for learning: Outcomes, practices, impact, and employers' views, 2011. Retrieved Jan 19, 2017 from <https://www.aacu.org/publications-research/publications/leap-vision-learning-outcomes-practices-impact-and-employers>.
- [20] A. of American Colleges Universities. E-portfolios. Retrieved on Jan 19, 2017 from <https://www.aacu.org/eportfolios>.
- [21] S. E. Page. *Diversity and Complexity*. Primers in Complex Systems. Princeton University Press, 2010.
- [22] Princeton Office of the Dean of the College. Grading at Princeton, 2015. Retrieved on Jan. 19, 2017 from <http://odoc.princeton.edu/faculty-staff/grading-princeton>.
- [23] J. Schinske and K. Tanner. Teaching more by grading less (or differently). *CBE-Life Sciences Education*, 13(2):159–166, 2014.
- [24] G. Siemens and P. Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.
- [25] J. W. Young. Adjusting the cumulative gpa using item response theory. *Journal of Educational Measurement*, pages 175–186, 1990.