WHO TAKES STATS IN US HIGH SCHOOLS? BACKGROUNDS, INTERESTS, & ASPIRATIONS

Emma M. Klugman, Gerhard Sonnert and Philip Sadler
Harvard University, USA
eklugman@g.harvard.edu

Statistics skills are increasingly required for a wide range of careers, and Statistics courses and degrees have exploded in popularity in recent years. We estimate that 920,000 US students are now taking Statistics classes in high school each year. We present results from a nationally representative survey of 15,727 college first-years attending two- and four-year institutions, of whom 26% had taken Statistics while in high school. We are the first to describe in detail this population of US high school Statistics course-takers, and present data about the demographics, career interests and values, STEM identity, grades, and test scores of those who took Statistics in high school. Latent profile analysis is used to characterize the profiles of key subgroups, illustrating the diverse skills, interests, and values of this population.

INTRODUCTION

Statistics courses have taken off at the high-school level in the United States. The number of students taking Advanced Placement (AP) Statistics exams more than doubled in the last decade: indeed, there were more AP Statistics exam-takers in 2019 (219,000 students) than for any of Chemistry, Spanish, Computer Science, Microeconomics, or Macroeconomics (College Board, 2021). Informally, when we have polled dozens of education researchers about how many students they estimate take the AP Statistics exam each year, even providing other subjects' totals as reference numbers, their guesses are around a tenth of this reality.

In terms of total students taking any Statistics course in high school, the number of AP exam-takers is clearly an underestimate. A fairer estimate is that 25% (NAEP, 2019) of US high-school students take Statistics before they graduate, of a projected 3.68 million students graduating high school each year (*Digest of Education Statistics*, 2020), giving an estimated 920,000 students taking Statistics in US high schools each year, a population about whom we know relatively little.

In this descriptive study, we profile the set of students taking Statistics in US high schools (who later proceeded to college), with two aims: (1) to simply study the demographics and correlates of Statistics-taking, and (2), to describe the different profiles of students taking Statistics in high school to uncover the diversity of interests and aspirations in this group. The data we use comes from our 2017 nationally-representative survey of college first-years, of whom 26% took a Statistics class in high school and receive our close attention.

DATA

Data for the present study were taken from our 2017 survey of 15,727 college students in the United States, funded by the National Science Foundation, grant number 1612375: *How Pre-College Informal Activities Influence Female Participation in STEM Careers.* The survey was administered to a nationally-representative sample of students enrolled in two- and four- year colleges in the United States, with a total of 119 institutions participating. A pencil-and-paper survey was administered to all first-year students in each institution, regardless of major. The nature of our dataset means we are unable to study high-school Statistics-takers who choose not to proceed to college, though we hope this important population will also receive further study.

BACKGROUNDS OF STUDENTS TAKING VARIOUS MATHEMATICS COURSES

Students were asked "Which of the following mathematics courses did you take in high school? *Mark all that apply*. Trig/Analytic Geometry, Pre-Calculus, Calculus, AP Calculus AB, AP Calculus BC, Statistics, AP Statistics, Integrated Math". We add two calculated fields here for "Any Statistics" and "Any Calculus". Table 1 shows a cross-tabulation of these results, along which the diagonal can be interpreted as the percentage of students taking each class (e.g., 19.2% of students took Statistics and 11.5% of students took AP Stat) and the off-diagonals showing the percentage of students taking both courses (e.g., 9.9% of students took both some form of Statistics and some form of Calculus).

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. https://iase-web.org

Table 1. Mathematics course taking in high-school, cross-tabulated

	Stat	APStat	(AnyStat)	Calc	APCalcAB	APCalcBC	(AnyCalc)	Trig	PreCalc	IntMath
Stat	19.2									
APStat	4.7	11.5								
(AnyStat)	19.2	11.5	25.9							
Calc	5.2	2.8	6.1	19.3						
APCalcAB	2.9	4.4	5.2	7	18					
APCalcBC	1.4	2.1	2.5	2.9	4.1	6.4				
(AnyCalc)	6.6	6.1	9.9	19.3	18	6.4	31.9			
Trig	12.5	7.2	16.3	13.6	11.9	4	21.3	53.7		
PreCalc	11.1	9	16.2	16.4	15.9	5.5	27.1	35.2	56.5	
IntMath	4.1	1.4	4.7	2.9	1.3	0.7	3.5	9.9	6.6	20.8

Several observations can already be made from these cross-tabulations. For example, the percentage of college first-years who report taking any Statistics class (25.9%) is not very far below the percentage who took some sort of Calculus class (31.9%). The percentage of students taking some sort of Statistics of these students who proceeded to college (25.9%) is also remarkably similar to the percentage of all high school 12th graders who've taken Statistics while in high school, according to NAEP data (25%), which suggests that those who do and don't attend college may take high school Statistics at similar rates (NAEP, 2019).

Of those students who took AP Calculus AB (18%), almost a third of them also took a Statistics class (5.2/18 ≈ 29%); for AP Calculus BC takers (6.4%), nearly 40% (2.5/6.4 ≈ 39%) also took a Statistics class. More than half of those who took AP Statistics (11.5%) also took some sort of Calculus class (6.1/11.5 ≈ 53%).

The modal college first-year student took neither Statistics nor Calculus. When we group students into four, mutually exclusive groups, of "Statistics" (only), "Calculus" (only), "Both", and "Neither", we see percentages of 16.1% Statistics, 22.1% Calculus, 9.9% Both, and 51.9% Neither, and we can compare these four groups on a range of demographic variables, as in Table 2 below. The size of our dataset means that most statistical tests for differences on this table show *p*-values below 0.001. We do not present these, as we prefer to focus on the magnitude of observed differences.

Table 2. Comparing the demographic characteristics of mutually exclusive course-taking groups

	Neither	Stat	Calc	Both
n	8172	2529	3474	1550
College Type				
Two-Year (%)	4076 (49.9)	869 (34.4)	563 (16.2)	236 (15.2)
High School Type				
HS_Public = TRUE (%)	6876 (86.9)	2141 (84.9)	2778 (80.2)	1137 (73.6)
HS_Private = TRUE (%)	585 (7.4)	260 (10.3)	431 (12.4)	310 (20.1)
Parents' Highest Education (%)				
<HS	566 (8.4)	151 (6.7)	165 (5.3)	55 (4.0)
HS	1249 (18.6)	363 (16.2)	349 (11.2)	127 (9.1)
AA	1831 (27.2)	515 (23.0)	519 (16.6)	187 (13.4)
BA	1708 (25.4)	656 (29.3)	984 (31.5)	413 (29.7)
>=MA	1378 (20.5)	554 (24.7)	1109 (35.5)	610 (43.8)
Gender (%)				
Male	3134 (42.6)	996 (41.6)	1608 (48.6)	763 (51.8)
Female	4170 (56.6)	1376 (57.5)	1683 (50.9)	702 (47.7)
Other	57 (0.8)	21 (0.9)	17 (0.5)	7 (0.5)
Language Primarily Spoken At Home				
English = TRUE (%)	5712 (79.0)	1863 (78.9)	2429 (74.4)	941 (65.0)
Racial Identity				
Hispanic = TRUE (%)	2078 (28.8)	552 (23.6)	599 (18.4)	163 (11.3)
Black = TRUE (%)	1159 (17.8)	323 (14.8)	280 (9.0)	112 (7.9)
White = TRUE (%)	4392 (67.4)	1517 (69.4)	2015 (65.1)	784 (55.3)
Asian = TRUE (%)	639 (9.8)	310 (14.2)	821 (26.5)	520 (36.6)
AmIn = TRUE (%)	236 (3.6)	52 (2.4)	64 (2.1)	33 (2.3)
OtherRace = TRUE (%)	613 (9.4)	144 (6.6)	154 (5.0)	71 (5.0)

We see that among our sample of college-goers, those 8,172 who took neither Statistics nor Calculus in high-school attended two- and four-year colleges at similar rates, whereas of the 1,550 students who had taken both Calculus and Statistics, only 15.2% were two-year college enrolees, the remaining all attending four-year colleges. Those who had taken Statistics only were more likely to have come from a public high-school than those who had taken Calculus, and less likely to have attended a private high-school. We see that students taking both Statistics and Calculus generally came from very highly educated homes, with parents holding high rates of graduate degrees, quite similarly to those who took Calculus only, whereas students taking only Statistics, or neither Statistics nor Calculus, were more likely to come from households with fewer degrees held. Statistics-only takers were more likely to be female than male, while both Calculus-only takers "Both" takers were more evenly split in terms of gender. Compared to Calculus courses, Statistics courses seemed to attract higher rates of students who identified as Black and Hispanic.

In general, it seems from our data that Statistics courses serve a more diverse and more disadvantaged population than Calculus courses, a finding which presents exciting opportunities for diversifying pathways into STEM, as well as an additional responsibility to serve these students excellently.

STATISTICS-TAKERS' INTERESTS AND ASPIRATIONS

Our survey also included questions about students' interests and aspirations, as well as various others. We selected nine major categories to focus on, and the summary statistics, as well as item-text, for each is presented in Table 3. These are the variables we later use to build our latent profiles.

Table 3 Summary statistics for selected interest and aspiration variables for Statistics-takers

Predictor:	Mean**:**	SD**:**	Measurement**:**
Interest in Math	2.67	1.72	"At the end of middle school how interested were you in Mathematics: Not at all interested 0, 1, 2, 3, 4, 5 Extremely interested"
Making Money	4.06	0.98	Career values: "Rate the following factors in terms of their importance for your future career satisfaction: - Making Money - Helping People Not at all important 0, 1, 2, 3, 4, 5 Extremely important"
Helping People	4.01	1.17	
HS ELA Grade	3.54	0.66	Students were asked, "What grade did you get in your last high school English course?" and selected from A, B, C, D, and F. We coded these as 4, 3, 2, 1, and 0.
SAT Score	1195	189	Participants were asked to self-report their standardized admissions test scores. SAT scores were asked in terms of buckets with ranges of 100 points (710-800, 610-700, etc.) for each subtest then summed. ACT scores were collected similarly then converted using concordance tables (*Guide to the 2018 ACT®/SAT® Concordance*, 2018).
STEM Identity	2.16	1.64	We used seventeen 0-5 scale Likert-type items asking the extent to which students agreed with statements such as "I enjoy learning about STEM", and "I see myself as a STEM person". Exploratory factor analysis suggested that these items were best understood as a single factor representing our hypothesized latent construct, so we took the simple mean of these scores as our "STEM identity" variable for ease of interpretation.
Legal, Social, and Cultural Professionals	0.36	0.67	Reported Career Interests: From a list of 24 different professions, students were asked to "mark all that apply" for which they had wanted to be as of the beginning of high school. Using the United Nation's International Standard Classification of Occupations, we group similar professions into three major categories:

Science and Engineering Professionals	0.41	0.81	- Legal Social and Cultural Professionals (*Lawyer, Anthro/Archaeologist, Social scientist, Humanities professional, Visual artist, Performing artist.)* - Science and Engineering Professionals (*Astronomer, Biologist, Chemist, Earth/Environmental scientist, Physicist, Other scientist, Engineer, CS/Programmer/IT Specialist, Math/Statistician.)*
Health Professionals	0.30	0.51	- Health Professionals (*Medical doctor, Health professional.)*

(International Labour Organization, 2008).
Within these groups, we simply took the sum of the number of professions each student expressed interest in (the means in this table represent sums, not proportions).

LATENT PROFILE ANALYSIS

In order to both characterize the demographics and correlates of Statistics-taking and to describe the different profiles of students taking Statistics in high school, we use latent profile analysis, a person-centred approach that groups students with similar profiles into distinct types (profiles). Readers unfamiliar with Latent Profile Analysis (LPA, also known as Normal Mixture Models) may find it helpful to conceive of the method as analogous to other cluster analysis methods with which they may be familiar, for example, *k*-means clustering or hierarchical clustering. One key difference is that LPA estimates the probabilities of profile membership for each student. Following Häfner et al. (2018) and Vermunt and Magidson (2003), we assume conditional independence for the sake of parsimony, that is, the correlations between observed characteristics were assumed to be fully explained by the latent profiles and no residual correlations were permitted.

We attempted to fit latent profile models for all high school Statistics-takers (n = 4,079) with each of one to eight profiles, but the solutions with seven and eight profiles failed to converge. The fit statistics for the first six solutions are shown in Table 4 below.

Table 4. Model Fit Indices for Solutions with 1-6 Profiles

Classes	LogLik	AIC	BIC	Entropy	n_min	n_max	BLRT_p
1	-52086.15	104208.30	104321.95	1.00	1.00	1.00	
2	-49413.78	98883.56	99060.34	1.00	0.27	0.73	0.01
3	-48977.85	98031.69	98271.61	0.80	0.27	0.45	0.01
4	**-48440.50**	**96976.99**	**97280.04**	**0.85**	**0.01**	**0.45**	**0.01**
5	-48604.78	97325.57	97691.75	0.75	0.08	0.27	0.01
6	-48597.83	97331.66	97760.98	0.73	0.04	0.28	0.01

Notes: LogLik = Log Likelihood, higher values (closer to zero) indicate better fit; AIC = Akaike information criterion, smaller values indicate better fit, favours parsimony over complexity; BIC = Bayesian information criterion, smaller values indicate better fit, also favours parsimony over complexity; Entropy represents how well the posterior probabilities from the model are able to confidently classify individuals, higher values are preferred; n_min = proportion of sample estimated to belong to the smallest profile; n_max = proportion of sample estimated to the largest profile; BLRT_p = p-value for the Bootstrapped-Likelihood Ratio Test, which compares the fit of an n profile solution to an n - 1 profile solution, and has been found to outperform the similar VLMR-LRT test in simulations, with a low p-value indicating that the larger model is preferred (Vermunt & Magidson, 2003, Nylund et al., 2007).

The Log-Likelihood, AIC, and BIC all point to the four-class solution as being optimal. It is peculiar that the two-class solution had an apparently perfect Entropy, but the four-class solution also does well on this metric. Though the BLRT indicates that ever more complex solutions are preferable, we choose to go with the weight of evidence that suggests that the four-class solution is best.

Figure 1, below, displays the selected predictors along the *x*-axis, and the standardized values of these along the *y*-axis. The four classes are each depicted using different colours and symbols shown in the key on the right-hand side. Looking at this figure, we can see that some variables, like "Making

Money" are not particularly helpful in differentiating between the four classes, while others, like math interest, STEM identity, and health career interest differ across the classes.

Figure 1. Four-class solution on standardized variables.

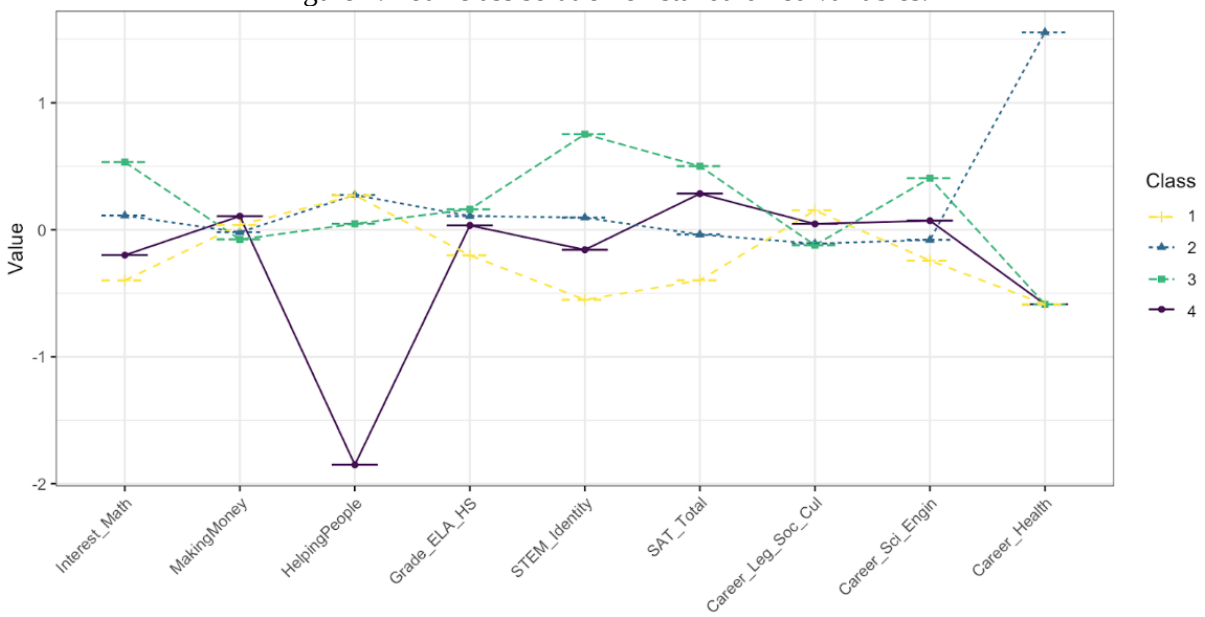

Table 5, below, shows the un-standardized means of each of the variables for the four different classes so we can meaningfully interpret them on their original scales. We selectively highlight boxes that we feel are helpful in characterizing each profile and order the four profiles from largest to smallest.

Table 5. Un-standardized means of the variables in each of the four classes.

Class	Percent	Money	Helping	Math_Int	ELA_Grade	SAT_Score	STEM_Iden	Leg_Soc_Cul	Sci_Engin	Health
1	38.66	4.11	4.32	1.97	3.41	1117	1.17	0.47	0.21	0.00
2	27.41	4.04	4.33	2.86	3.61	1188	2.32	0.29	0.35	1.09
3	25.13	3.97	4.05	3.70	3.67	1303	3.62	0.26	0.79	0.00
4	8.80	4.18	1.53	2.25	3.55	1250	1.87	0.41	0.47	0.00

The largest class, Class 1, is best characterized by having a high desire to help people through their careers, the lowest math interest and STEM identity, and the most interest in Legal, Social, and Cultural Careers. The next class, Class 2, has the highest desire to help people, and is otherwise around the middle of most other variables, except that this group has a much higher interest in health and medical careers. The third class has a slightly lower interest in making money, the highest math interest, ELA grade, SAT score, and STEM identity, and have the highest interest in science and engineering careers. Finally, the fourth, smallest class is somewhat confusing to us. They have the highest desire to make money, the lowest desire to help people, and quite low math interest and STEM identity.

DISCUSSION

Our study shows that students taking Statistics in US high schools are diverse in their skills, interests, and aspirations. For most of these students, this is their first exposure to our field, and for many, it will also be their last. Better understanding these high school statistics takers will hopefully, in the long-run, allow us to (1) design better experiences that increase the likelihood that students take further courses in Statistics and Data Science, (2) improve student attitudes toward Statistics, and (3) reduce inequities in access to Statistics and Data Science, and data-driven careers more broadly.

It is our hope that educators will find our analysis of common Statistics student profiles to be helpful in knowing their audience so that they can target and differentiate their teaching to appeal to students with these interests and backgrounds. For example, in the US high school advising context, calculus and statistics classes are often pitted against each other, but we find that many students take

both. Further, though a stereotypical calculus-taker might have strong mathematical preparation and a strong interest in STEM careers, this profile also describes a sizeable portion of statistics-takers. Studies like this one can help to counter stereotypes about student profiles with data, as well as draw attention to diversity within groups, rather than merely differences across groups.

Another finding we wish to highlight is the demographic information from Table 2, showing how statistics-takers, calculus-takers, both-takers, and neither-takers differ along several dimensions of privilege and representation. While many conclusions could be drawn from these comparisons, we wish to celebrate the teachers who work hard to make statistics accessible to this diverse population of students, including many disadvantaged students.

Our study suggests that policymakers and administrators should remember that Statistics is an important part of the pathway for many students: some who may only ever need or want one course, and others who will need more advanced training; some who are equipped for a rigorous, mathematical introduction, and others who are not; some who aspire to STEM and health professions, and others who prefer legal, social, and cultural professions. This should certainly inform course offerings, curricula, and design, to best meet the needs of these different students.

LIMITATIONS

The results of any Latent Profile Analysis depend on choices made by the researchers. Though we found that the different model specifications and different predictors we explored led to similar profile solutions, we acknowledge that other researchers may have selected different predictors and come up with different results than our own. A further limitation is that our data source only allows us to characterize high school Statistics-takers who ultimately proceed to college, and therefore misses the many students who take Statistics in high-school and then select other post-secondary pathways. This study is limited also by the fact that the survey data we used was not collected for this purpose of studying Statistics students. This means that our work was highly inferential: having asked college students about their activities in high school, we sought to understand their motivations for taking Statistics. Our hope is that this study inspires further research that ventures into the high school to talk to students as they make their first decisions about whether to study Statistics.

ACKNOWLEDGEMENTS

Project undertaken with support from NSF Grant # 1612375

REFERENCES

College Board. (2021). *AP Data: College Board*. AP Data - Research - College Board. https://research.collegeboard.org/programs/ap/data

Digest of Education Statistics. (2020). National Center for Education Statistics. https://nces.ed.gov/programs/digest/d20/tables/dt20_219.10.asp

Guide to the 2018 ACT®/SAT® Concordance (p. 8). (2018).

Häfner, I., Flunger, B., Dicke, A.-L., Gaspard, H., Brisson, B. M., Nagengast, B., & Trautwein, U. (2018). The Role of Family Characteristics for Students' Academic Outcomes: A Person-Centered Approach. *Child Development*, *89*(4), 1405–1422. https://doi.org/10.1111/cdev.12809

International Labour Organization. (2008). *ISCO - International Standard Classification of Occupations*. http://www.ilo.org/public/english/bureau/stat/isco/isco08/index.htm

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

NAEP. (2019). *National Assessment of Educational Progress—Mathematics, Grade 12, 2009, 2013, 2015, 2019*. National Assessment of Educational Progress.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Vermunt, J. K., & Magidson, J. (2003). Latent class models for classification. *Computational Statistics & Data Analysis*, 7.