Using Human-Centered Artificial Intelligence to Assess Personal Qualities in College Admissions

Benjamin Lira[1,*], Margo Gardner[2], Abigail Quirk[1], Cathlyn Stone[2], Arjun Rao[2], Lyle Ungar[1], Stephen Hutt[1], Sidney K. D'Mello[2,†], and Angela L. Duckworth[1,†]

[1]University of Pennsylvania; [2]University of Colorado-Boulder

This manuscript was compiled on January 20, 2023

There is mounting evidence that personal qualities predict an array of life outcomes, including success in college. Unfortunately, the holistic process by which prosocial purpose, leadership, and other personal qualities are considered in college admissions can be prohibitively resource-intensive. Could artificial intelligence (AI) improve this process? AI has been criticized as a "black box" approach that can inadvertently penalize already disadvantaged subgroups. As an alternative, we assess a Human-Centered Artificial Intelligence (HCAI) approach to assessing personal qualities from text. First, human raters coded 3,131 applicant essays describing out-of-school extracurricular and work experiences for seven different personal qualities. Next, a pre-trained language model fine-tuned on this data successfully reproduced human codes (average *r* = 0.72) and did so equally well across demographic subgroups. Finally, in a larger, national sample (*N* = 309,594), computer-generated scores collectively demonstrated incremental predictive validity for six-year college graduation. Taken together, our findings highlight both challenges and opportunities of HCAI for the efficient, equitable, and interpretable assessment of personal qualities.

Many colleges embrace the ideals of holistic review. In a recent survey by the National Association for College Admissions Counseling, 70% of admissions officers said they consider personal qualities to be an important factor when selecting applicants (1). This aim is justified by longitudinal research affirming that personal qualities—whether referred to as "non-cognitive skills," "social-emotional competencies," "personality," or "character"—predict positive life outcomes in general and success in college in particular (2–4). Moreover, it has been argued that a holistic admissions process advances equity insofar as applicants are able to demonstrate qualifications not reflected in their standardized test scores, which tend to be highly correlated with socioeconomic advantage (5).

However, history suggests that equity is by no means guaranteed by holistic review. A century ago, Columbia University first began requiring applicants to write a personal essay, which admissions officers evaluated for evidence of "good character" (6). Previously, the university's admissions decisions had been based primarily on standardized test scores. The result was a growing proportion of Jewish students in each entering class, which in turn led to concerns that, as Columbia's dean at the time put it, the campus was no longer welcoming to "students who come from homes of refinement" (p. 87). It has been argued that because summary judgments of character were in the eye of the beholder (i.e., the admissions officer), this more holistic review process made it possible to unfairly penalize Jewish applicants.

Although its aims may be nobler today, the holistic review process itself remains much the same. College admissions officers still rely on the personal essay to evaluate an applicant's personal qualities (1). Best practices recommend that when doing so, they assess personal qualities separately (as opposed to making a summary judgment of "good character"), use a structured rubric (as opposed to intuition), and carry out these evaluations independently of one another (as opposed to relying on a single officer's judgment) (5, 7). Such practices represent the application of basic psychometric principles and, in the research context, have long been used to increase the reliability, validity, and interpretability of human ratings (8, 9). Moreover, the transparency of this systematic approach should limit bias—whether accidental or intentional. In college admissions, however, this ideal is hardly ever achieved. The soaring number of applications that admissions officers must review—which for the majority of colleges has more than doubled in the last two decades—affords extraordinarily limited time to review each one (10, 11). In sum, logistical and budgetary realities prohibit the implementation of procedures that, if resources were unlimited, could optimize reliability, validity, and interpretability, and in turn, equity.

Should artificial intelligence (AI) be used to advance the aims of holistic review? With stunning efficiency, AI systems are able to identify patterns in data and, with stunning fidelity, to apply learned models to new cases. Once trained, a computer algorithm could generate personal quality scores from student writing instantaneously, reliably, and at near-zero marginal cost. However, there are concerns that the "black box" of an AI algorithm would inadvertently perpetuate, or even exacerbate, bias against disadvantaged subgroups. This has been shown in the domains of hiring, criminal justice, and medical diagnosis (12–14). In college admissions, AI-quantified essay content and style have been shown to correlate more strongly with household income than do SAT scores (15). Opaque AI algorithms that provide fertile ground for bias recall the anti-semitic holistic review practices of a century ago. Arguably, by allowing bias to remain hidden, any "black box" process can undermine, rather than advance, equitable decision making.

In this investigation, we evaluated a Human-Centered Artificial Intelligence (HCAI) approach to assessing personal qualities. HCAI is an emerging approach to AI that prioritizes equity and interpretability (16), making automation a complement rather than a substitute for human control (17). In contrast to conventional AI approaches, HCAI begins with priorities articulated by human stakeholders, then builds mod-

[†]SKD and ALD share senior authorship of this project

[*]To whom correspondence should be addressed. E-mail: blira@upenn.edu

Table 1. Personal qualities and example essay excerpts

Personal quality	Fictionalized excerpts
Prosocial purpose Helping others, wanting to help others, consideration of the benefits to others, mention of reasons for helping others, or reflection on how enjoyable or rewarding it is to help others.	Every summer for the last three years, I worked as camp counselor at a camp for young children from underprivileged families. Helping children realize their hidden talents is one of the most rewarding experiences I have ever had. I've been so fulfilled by watching these children develop confidence in their abilities. This experience has been so important to me, and it showed me that a career in education is where I belong.
Leadership Serving in a leadership role, commenting on what he or she did in his or her capacity as a leader, or discuss the value, meaning, or importance of leadership.	I was chosen to be cheerleading captain during my senior year. My freshman year captain had a huge impact on my life, and I felt like it was my time to pay it forward. I am so proud of everything I did for the girls: creating a mentorship system, organizing events and fundraisers, and encouraging everyone to work as hard as they could. At the end of the year, a few girls thanked me. I was completely overcome with emotion. I've never felt so gratified in my life.
Learning Improving, learning, or developing knowledge, skills, or abilities.	I played softball in high school. When I started, I was not a very strong player. When I finally made the varsity team my senior year, I was determined to have a better season. I worked constantly to improve my game – during practice and on my own time. My skills grew so much. Because of my hard work, I finished the year with the best record on my team!
Goal pursuit Having a goal and/or a plan.	I have been playing soccer since I was six years old. Unfortunately, last year I injured my knee, and it has been a struggle to get back to the level I was playing at before my injury. It has been really challenging, but I've been doing physical therapy and practicing everyday so that I can be a varsity starter this year.
Intrinsic motivation Describing the activity as enjoyable or interesting. Liking the activity or identifying with it.	Running track is so much more than a sport to me. It's a challenge and an adventure, and I put everything I have into it. I love every aspect of it, even the afternoons I spend drenched in sweat in the scorching heat.
Teamwork Working with or learning from others. Valuing what fellow participants bring to the activity.	I've been on my school's debate team since my freshman year, and was elected co-captain because of my commitment to the team's success. My fellow co-captains and I worked together to get our team ready for competitions. We knew that a strong team performance was more important than the successes of a few individuals. We stressed teamwork and cooperation between our teammates. Because we focused on team effort, we earned first place at the state meet.
Perseverance Persisting in the face of challenge.	I've learned to become a gracious victor and to grow from defeat. Track has helped me overcome my fear of losing, and even helped me put my life in perspective. I've learned to keep working and fighting even when the odds seem impossible to beat. There were many times that I found myself lagging, but I pulled ahead at the end because I never gave up. The most important thing I've learned is to never let anything stand in my way.

els whose internal logic is interpretable by those stakeholders, and finally, audits computer model outputs for unintended bias. We began with a de-identified sample of 309,594 college applications (see **Figure 1**), each of which included a 150-word essay describing an extracurricular or work activity of the applicant's choice. Next, in a *Development Sample* of 3,131 essays, human research assistants to identify seven different personal qualities commonly valued by universities and shown in prior research to predict college success (3). See **Table 1**. These ratings were used to fine-tune the RoBERTa language model (18) for each personal quality. We then confirmed each model's interpretability, as well as evidence of its convergent, discriminant, and predictive validity by demographic subgroup. Finally, we applied these fine-tuned models to the *Holdout Sample* of 306,463 essays, examining associations between computer-generated personal quality scores and demographic characteristics and six-year college graduation.

Results

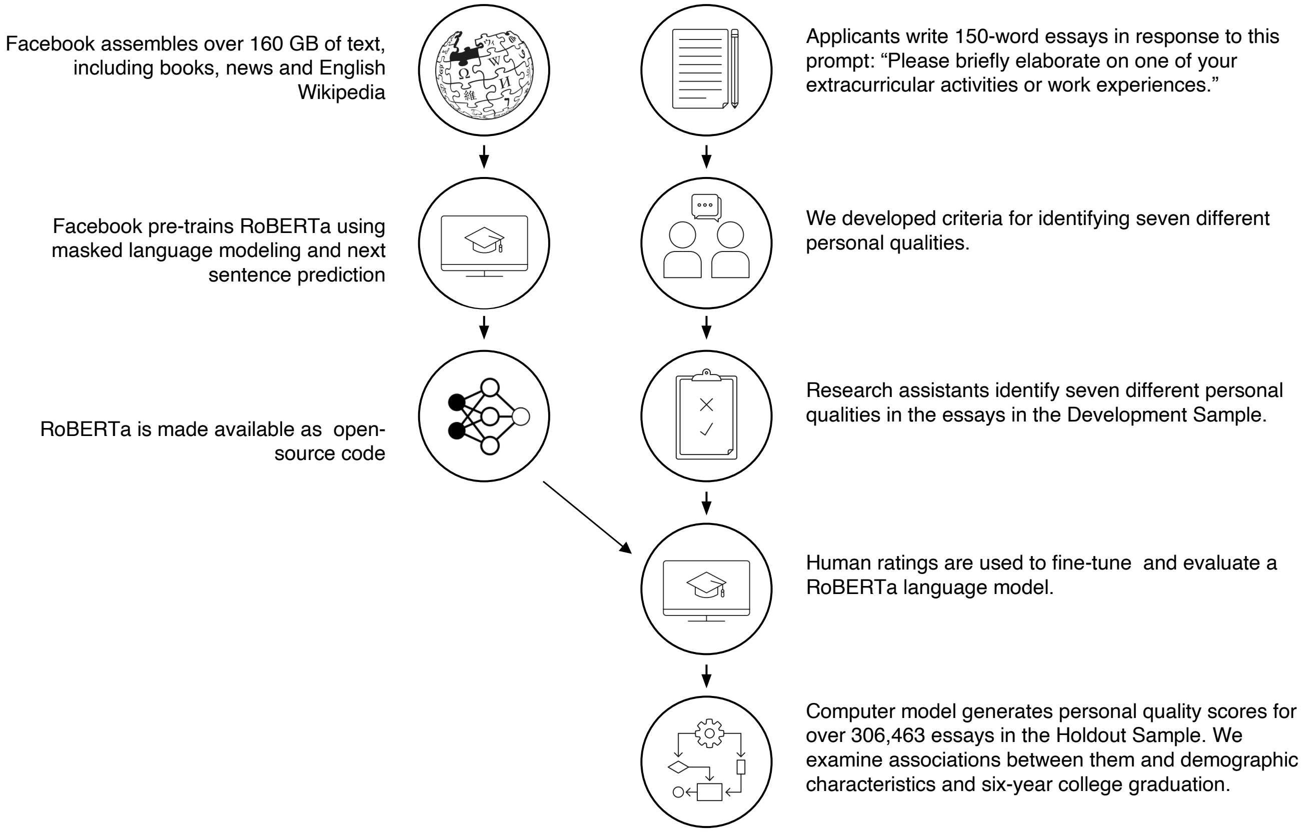
On average, human raters found evidence for two personal qualities in each essay. As shown in Table 2, some personal qualities were more commonly observed than others. For instance, raters identified the personal qualities of learning and intrinsic motivation in 42% of essays, whereas they identified leadership and perseverance in only 18% and 19% of essays, respectively.

Using these binary human ratings, we fine-tuned separate RoBERTa models to produce continuous likelihood scores for each personal quality. See **Supplementary Materials** for details.

Model interpretability. We used the *transformers-interpret* package (19, 20) to identify the words (or fractions of words) that these fine-tuned RoBERTa models relied on most to generate personal quality scores. As shown in Figure 2, there was reasonable evidence of face validity. For instance, RoBERTa assigned higher scores for leadership when essays mentioned "president," "leader," and "captain." See **Supplementary Materials** for details.

Convergent and discriminant validity of computer-generated likelihoods in the Development Sample. Computer-generated likelihoods for each personal quality converged with human ratings of the same personal quality (rs ranged from .59 to .86, average r = .74). In contrast, computer-generated likelihoods for a particular personal quality did not correlate with human ratings of other personal qualities (rs from -.16 to .18, average r = .01). See **Table 2.** Not surprisingly, the more reliably human raters were able to code each personal quality, the better the computer-generated likelihoods of personal qualities matched these ratings ($r = .95$, $p = .001$). In the sub-sample of essays that were coded by multiple raters, model scores agreed more strongly with human ratings than human ratings agreed with each other ($\mathsf{M}_{\mathsf{human-computer}} = .73$, $\mathsf{M}_{\mathsf{human-human}} = .69$, t = 4.16, p = .006).

Convergent validity does not vary by demographic subgroup in the Development Sample. As shown in **Table 3,** correlations



Facebook assembles over 160 GB of text, including books, news and English Wikipedia

Facebook pre-trains RoBERTa using masked language modeling and next sentence prediction

RoBERTa is made available as open-source code

Applicants write 150-word essays in response to this prompt: “Please briefly elaborate on one of your extracurricular activities or work experiences.”

We developed criteria for identifying seven different personal qualities.

Research assistants identify seven different personal qualities in the essays in the Development Sample.

Human ratings are used to fine-tune and evaluate a RoBERTa language model.

Computer model generates personal quality scores for over 306,463 essays in the Holdout Sample. We examine associations between them and demographic characteristics and six-year college graduation.

Fig. 1. A Human-Centered Artificial Intelligence (HCAI) approach to assessing personal qualities in college admissions.

Table 2. Descriptive statistics and correlations between human ratings and computer-generated likelihoods of personal qualities in the Development Sample

	Human Ratings						
Personal Quality	PP	LD	TW	LR	PS	IM	GP
Computer-Generated Likelihoods							
1. Prosocial purpose (PP)	.86***	−.01	−.04*	−.09***	−.12***	−.05**	.04*
2. Leadership (LD)	−.01	.81***	.15***	−.01	.00	−.09***	.05**
3. Teamwork (TW)	−.07***	.18***	.62***	.06**	.07***	−.02	.06**
4. Learning (LR)	−.10***	−.05**	.04*	.77***	.11***	−.01	−.03
5. Perseverance (PS)	−.16***	−.01	.06**	.10***	.67***	.03	.05**
6. Intrinsic motivation (IM)	−.05**	−.09***	.00	−.03	.04*	.73***	.03
7. Goal pursuit (GP)	.06***	.06**	.06***	−.01	.02	.02	.59***
Descriptive Statistics							
Human Inter-Rater Reliability	.83	.78	.61	.73	.66	.63	.57
Frequency of Human Rating	.34	.18	.26	.42	.19	.42	.31
Mean of Computer-Generated Likelihood	.36	.19	.26	.45	.19	.45	.32

Note. Inter-rater reliability for human raters was measured with Krippendorf’s alpha. Correlations between human ratings and computer-generated likelihoods for the same personal qualities are shown in bold. All correlations are point-biserial correlation coefficients between binary human ratings and continuous computer-generated likelihoods. * $p < .05$. ** $p < .01$. *** $p < .001$. $n = 3{,}131$. n for inter-rater reliability = 206 essays coded by multiple raters.

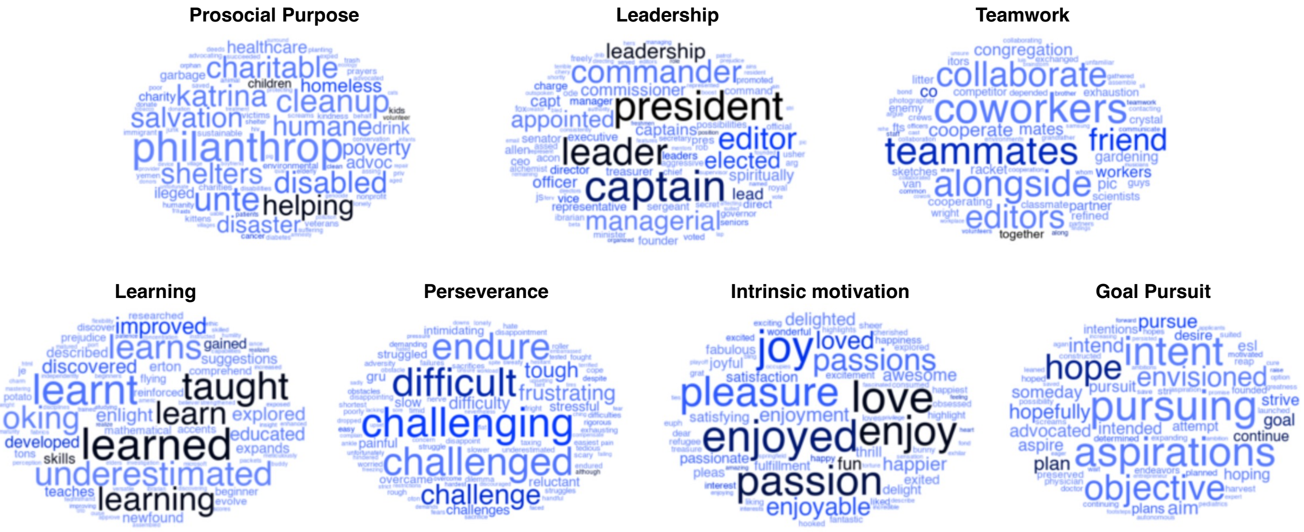


Fig. 2. Complete or partial words on which fine-tuned RoBERTa models relied most for generating personal quality scores. Font size is proportional to word importance. Darker words are more common. Token "gru" is a fraction of the word "grueling", Token "unte" is a fraction of the word "volunteer". Words importance is not invariant across essays, it depends on word context. Word importance and frequency were largely independent (r = -.03, p < .001). For instance, for intrinsic motivation, the model relied more on the word "pleasure" then the word "fun," but essays were more likely to contain the word "fun" then the word "pleasure."

Table 3. Correlations between human ratings and computer-generated likelihoods of personal qualities by demographic subgroup in the Development Sample

Demographic Category	*n*	PP	LD	TW	LR	PS	IM	GP	ACV	ADV	Range of DV	
Race/Ethnicity												
White	871	0.87	0.79	0.59	0.80	0.68	0.73	0.58	0.74	−0.01	−0.15	0.14
Black	487	0.84	0.78	0.64	0.76	0.72	0.78	0.60	0.74	0.00	−0.27	0.22
Latino	501	0.86	0.85	0.60	0.78	0.63	0.71	0.67	0.74	0.02	−0.14	0.21
Asian	590	0.84	0.80	0.62	0.72	0.67	0.70	0.54	0.71	0.01	−0.15	0.20
Other	290	0.84	0.83	0.61	0.72	0.56	0.74	0.54	0.71	0.01	−0.21	0.23
No race reported	369	0.90	0.85	0.66	0.77	0.70	0.75	0.65	0.77	0.01	−0.14	0.20
Parents with college degrees												
None	1,608	0.85	0.81	0.60	0.78	0.66	0.75	0.61	0.74	0.01	−0.19	0.20
One	653	0.86	0.81	0.62	0.79	0.65	0.71	0.57	0.73	0.01	−0.12	0.17
Two	853	0.88	0.80	0.63	0.74	0.69	0.71	0.58	0.73	0.00	−0.14	0.19
Gender												
Female	1,702	0.86	0.81	0.62	0.77	0.68	0.73	0.60	0.74	0.00	−0.18	0.18
Male	1,413	0.85	0.81	0.61	0.77	0.66	0.74	0.58	0.73	0.00	−0.15	0.17
Married parents												
Parents married	2,055	0.85	0.81	0.61	0.77	0.68	0.74	0.58	0.73	0.00	−0.15	0.17
Parents not married	1,061	0.88	0.80	0.63	0.77	0.66	0.71	0.62	0.74	0.01	−0.18	0.19
English language learner status												
English language learner	808	0.87	0.77	0.59	0.74	0.69	0.72	0.61	0.73	0.01	−0.15	0.17
Native speaker	2,308	0.86	0.82	0.62	0.78	0.66	0.73	0.59	0.74	0.00	−0.17	0.18
Title 1 status of high school												
Title 1 public school	1,127	0.83	0.84	0.61	0.77	0.67	0.74	0.59	0.74	0.01	−0.21	0.21
Non-Title 1 school	1,552	0.88	0.80	0.63	0.78	0.66	0.72	0.60	0.74	0.00	−0.12	0.16

Note. All correlations are point-biserial correlation coefficients between binary human ratings and continuous computer-generated likelihoods. All correlations were significantly different from zero (p < .001). ACV (average convergent validities) are the average correlations between human ratings and computer-generated likelihoods for the same personal qualities. ADV (average discriminant validities) are the average correlations between human ratings and computer-generated likelihoods for differing personal qualities. n = 3,131.

between human ratings and computer-generated likelihoods of personal qualities were similar across subgroups. For example, the average correlation between human-rated and computer-generated personal quality scores was .74 for female applicants and .73 for male applicants. As shown in **Table S6,** After correcting for multiple comparisons, 13% of the correlations differed by subgroup. In about half of these comparisons, the models were more accurate for the marginalized group, while in the other half, the majority subgroup was favored. In most of these cases the difference between the correlations was not very large (mean $|\Delta r| = .054$, range of Δr = -.121 – .056).

Human ratings and computer-generated likelihoods were largely unrelated to demographics in the Development Sample. Demographic characteristics were largely unrelated to personal qualities whether assessed by human raters (mean $|\phi| = 0.02$) or by computer algorithm (mean $|d| = 0.06$). One exception is that female applicants were rated as more prosocial than male applicants ($\phi = 0.13$, $p < .001$ for human ratings, $d = 0.26$, $p < .001$ for computer-generated likelihoods, p-values adjusted for multiple comparisons)—in line with other research showing gender differences in prosocial motivation and behavior favoring women (21). See **Table S3** in **Supplementary Materials** for details.

Computer-generated likelihoods were as predictive of college graduation as human raters in the Development Sample. To compare the predictive validity of the computer-generated likelihoods with human ratings, we ran two logistic regression models in which personal qualities predicted college graduation. The computer generated likelihoods were slightly more predictive than the human ratings, but the difference in the AUCs was not significant ($\mathsf{AUC}_{\mathsf{human}} = .565$, $\mathsf{AUC}_{\mathsf{computer}} = .574$, $\Delta\mathsf{AUC} = .009$, $p = .274$). Coefficients were significantly larger for computer-generated likelihoods as opposed to human ratings ($t = 2.99$, $d = 1.13$, $p = .024$).

Computer-generated likelihoods were largely independent of demographics but, in support of criterion validity, predicted graduation in the Holdout Sample. Next, we applied the fine-tuned models to the *Holdout Sample* of 306,463 essays. For each personal quality, reliability across models trained on different subsets of the data was high (range of Cronbach's α = .990 to .997). Even when considering any two models, they were likely to produce similar results (average inter-model correlation ranged from .910 to .967).

As in the development sample, computer-generated likelihoods for personal qualities were similar across demographic subgroups (mean $|d| = 0.05$). In contrast, and as expected, demographics were more strongly related to standardized test scores (mean $|d| = 0.38$) and degree of participation in out-of-school activities (mean $|d| = 0.17$). See **Supplementary Materials** for details.

About 78% of students in the *Holdout Sample* graduated from college within 6 years. As shown in Model 1 in **Table 4,** computer-generated likelihoods for personal qualities were each modestly predictive of college graduation when controlling for each other (*OR*s from 1.041 to 1.132, ps < .001; AUC = .560). To estimate a ceiling on how much the essays can predict subsequent college graduation, we trained a RoBERTa model to predict college graduation from students' responses. This model achieved an out-of-sample AUC of .626, suggesting

Table 4. Odds ratios from binary logistic regression models predicting six-year college graduation in the *N* = 306,463 Holdout Sample

	(1)	(2)
Computer-Generated Likelihoods of Personal Qualities		
Prosocial Purpose	1.132***	1.075***
	(0.005)	(0.005)
Leadership	1.133***	1.065***
	(0.005)	(0.005)
Teamwork	1.080***	1.031***
	(0.005)	(0.005)
Learning	1.065***	1.045***
	(0.004)	(0.005)
Perseverance	1.071***	1.012**
	(0.005)	(0.005)
Intrinsic Motivation	1.068***	1.007
	(0.004)	(0.005)
Goal Pursuit	1.041***	1.005
	(0.004)	(0.005)
Race/ethnicity (vs. White)		
Black		0.774***
		(0.019)
Latino		0.871***
		(0.019)
Asian		0.735***
		(0.017)
Other		0.749***
		(0.017)
No race reported		0.849***
		(0.013)
Parental education (vs. No parent w/ college degree)		
One parent w/ college degree		1.199***
		(0.012)
Two parents w/ college degree		1.335***
		(0.012)
Female		1.435***
		(0.010)
Married parents		1.311***
		(0.011)
English language learner		0.769***
		(0.015)
Title 1 high school		0.951***
		(0.013)
Out-of-school activities (OSA)		
Number of OSA		1.250***
		(0.005)
Time per OSA		1.088***
		(0.004)
Proportion sports		1.042***
		(0.005)
Standardized test scores		1.489***
		(0.006)
Constant	3.555***	2.533***
	(0.004)	(0.014)
AUC	.560	.689

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

that consistent with previous research (15) essays do encode information predictive of graduation outside of personal qualities. The same procedure using personal qualities results in an out-of-sample AUC of .557. See **Supplementary Materials** for details.

As shown in Model 2 in **Table 4,** five of seven personal qualities remained predictive of college graduation when controlling for each other, demographics, standardized test scores, and out-of-school activities (ORs from 1.012 to 1.075, ps $< .01$). See **Supplementary Materials** for details on imputation and robustness checks.

As a further test for fairness, we tested whether the predictive power of computer-generated likelihoods of personal qualities was equivalent across subgroups. We added interaction terms between each personal quality and standardized test scores and each demographic characteristic. After controlling for multiple comparisons (22), we confirmed that the predictive effect of personal qualities was equal across demographic subgroups. Comparatively, the predictive accuracy of standardized tests differed across subgroups (mean $|\beta| =$ -0.053). See **Supplementary Materials** for details.

Discussion

We evaluated a Human-Centered Artificial Intelligence approach to measuring personal qualities from student writing using a national dataset of over 300,000 college applications. Specifically, we fine-tuned RoBERTa language models using human ratings of prosocial purpose, leadership, teamwork, learning, perseverance, intrinsic motivation, and goal pursuit, respectively, in applicants' essays about their out-of-school activities. These models demonstrated convergent, discriminant, and predictive validity—and this evidence was consistent across demographic subgroups. In addition, computer-generated scores were largely independent of demographics.

In contrast, two prior studies found that AI-extracted admission essay content and style correlate with socioeconomic status. Alvero and colleagues (15) found that students from wealthier families tend to write about certain essay topics (e.g., human nature), whereas disadvantaged students tend to write about others (e.g., tutoring groups). Likewise, Pennebaker and colleagues (23) found that categorical words (e.g., articles, prepositions) versus dynamic words (e.g., pronouns, adverbs) in college essays correlate with parental education at $r = .22$. Why do our results differ? It seems likely that personal qualities are distributed more evenly across demographic subgroups than the topics students choose to write about or the words they use to do so. However, we cannot rule out methodological differences. Alvero et al. (15) used essays from the University of California system, and Pennebaker et al. (23) used essays from a large state university. In contrast, our sample included a larger and more diverse set of public and private four-year colleges from across the United States. In addition, both of these prior studies used personal statements totaling several hundred words, whereas the essays to which we had access were a maximum of 150 words and focused specifically on extracurricular activities and work experiences. Finally, rather than using unsupervised topic modeling or dictionary approaches, we fine-tuned a language representation model using human ratings.

Several limitations of this investigation suggest promising directions for future research. First, while our national dataset was unusually large and diverse, it did not include the 650-word personal essay now required by the Common Application. Unfortunately, applicants in 2008-09 submitted their personal essays as attached PDF files that were not feasible to de-identify. A replication and extension of our study using a more recent cohort of applicants should not face this limitation. Second, and relatedly, because the majority of applicants in our sample submitted their high school transcripts as attached PDF files that could not be de-identified, our dataset included high school GPAs for only a subsample of 43,592 applicants whose school counselors entered grades directly into the Common Application online portal. While our robustness check using this subsample (see **Supplementary Materials Table S25**) affirm the conclusions of our main analyses, future research should not face this limitation.

Third, the observed effect sizes for personal qualities predicting college graduation were modest, both in absolute terms and relative to the predictive validity of standardized test scores. Future research should include other important outcomes—including more proximal and continuous ones—such as college GPA, extracurricular involvement, mental health, and contributions to the campus community (24). On the other hand, a growing literature suggests that life outcomes are generally extremely difficult to predict with precision (25, 26), in part because the more factors there are that determine an outcome, the smaller the influence of any single one (27). Regardless, it seems likely that more important to graduation than personal qualities are factors we could not measure, including the student's ability to afford tuition payments (28), their academic preparation and support (29, 30), and their sense of belonging (31).

Finally, the inter-rater reliability of human raters was less than ideal for certain personal qualities (e.g., intrinsic motivation). Not surprisingly, the personal qualities that were more reliably coded (e.g., prosocial purpose) were more accurately detected by the algorithm and also more predictive of college graduation. In other words, consistent with the adage "garbage in, garbage out," the quality of an AI model depends on the quality of its training data (32). We speculate that as subject matter experts, college admissions officers might, both in future research and in practice, produce more reliable ratings than the trained research assistants in our investigation. Likewise, aggregating essays with other sources of information (e.g., letters of recommendation from teachers and guidance counselors) would no doubt strengthen reliability and, in turn, validity (cf.(2, 9, 33, 34)).

What are the practical implications of our results? Consistent with the core tenets of Human Centered AI, we recommend artificial intelligence be used to augment, not replace, human judgment. No algorithm can decide what the goals of a university's admissions process should be, nor what personal qualities matter most for that community. What's more, there will always be a need for humans to use common sense to verify algorithmic outputs. Algorithms look for patterns and, thus, by design are blind to exceptions. For instance, our fine-tuned RoBERTa model gives the sentence "I donated heroin to the children's shelter" an extremely high score for prosocial purpose. Seeing algorithms as complements rather than replacements for human judgment may also counter algorithm aversion—the tendency to trust human decision makers over algorithms, even in the face of contradictory evidence

(35).

In sum, this investigation suggests that a Human-Centered Artificial Intelligence approach to measuring personal qualities warrants both optimism and caution. Our findings demonstrate that AI models trained on human ratings are not only efficient (yielding millions of personal quality scores in a matter of minutes) and reliable (replicating human ratings with uncanny precision) but also interpretable (as opposed to an inscrutable "black box") and auditable (e.g., for fairness to demographic subgroups). Like any new technology, however, there may be unintended collateral consequences. Campbell's Law (36) states that the more weight given to an assessment in high-stakes decisions (as opposed to low-stakes research), the greater the incentive for distortion. It is not hard to imagine how applicants might try to mold their essays—perhaps using AI tools like chatGPT—to match what admissions officers, and the algorithms they train, are looking for. We can only assume that applicants from more advantaged backgrounds would be better positioned to do so. Nevertheless, progress in any field depends on dissatisfaction with the status quo, and there is no doubt that when it comes the assessment of personal qualities in college admissions, we can do better.

Materials and Methods

Participants. After exclusions, our sample consisted of 309,594 students who applied to universities in 2008-09. To provide labeled data for the machine learning algorithm, we set aside a *Development Sample* consisting of 3,131 applications for manual coding. We used stratified random sampling to ensure representation across demographic groups and levels of involvement in extracurricular activities. The *Holdout Sample* was composed of the remaining 306,463 essays. We applied the fine-tuned algorithm to these essays and tested the relationship between the computer-generated likelihoods of personal qualities and demographics as well as college graduation. See **Supplementary Materials** for details on missing data and exclusion criteria.

Measures.

Extracurriculars Essay. In up to 150 words, applicants who completed the Common Application were asked to respond to the following prompt: "Please briefly elaborate on one of your activities or work experiences." We excluded all essays shorter than 50 characters, most of which were mentions to attachments (e.g., "See attached").

Standardized test scores. Over half (55%) of the Holdout Sample reported SAT scores, 14% reported ACT scores, 25% reported both, and 6% reported neither. Using published guidelines, we converted ACT scores to SAT scores. For students who reported both test scores, we selected the higher score, and for students who reported neither, data were considered missing.

Extracurricular activities. Applicants listed up to seven extracurricular activities and for each, indicated the years they had participated. For each applicant, we computed the total number of extracurricular activities, mean years per activity, and the proportion of activities that were sports.

Demographics. We obtained the following demographic information from the Common Application: race/ethnicity, parental education, gender, parents' marital status, English language learner status, and type of high school (i.e., Title 1 public school vs. other kinds of schools).

College graduation. We obtained data from the 2015 National Student Clearinghouse (NSC) database (www.studentclearinghouse.org) to create a binary six-year graduation measure (0 = did not earn a bachelor's degree from a four-year institution within six years of initial enrollment; 1 = earned a bachelor's within six years). We obtained institutional rates of graduation within six years from the National Center for Educational Statistics (NCES). We control for any potential effects of baseline institutional effects on the odds of graduation in the **Supplementary Materials Table S26**.

Analytic Strategy. To handle missing data, we used multiple imputation ($m = 25$), employing the mice package in R (37). We used predictive mean matching for graduation rates and college admissions test scores. For school type, we used polytomous regression. In the *Holdout Sample,* 5.7%, 12.2%, and 7.1% of students were missing data on admissions test scores, six-year institutional graduation rates, and high school Title 1 status, respectively.

In binary logistic regression models we standardized all continuous variables to facilitate interpretation of odds ratios. Factor variables were dummy-coded and, along with binary variables, were not standardized, such that the effects shown indicate the expected change in the odds of each variable relative to the comparison group.

When averaging correlations together, we transformed the correlation coefficients to z-scores using Fisher's transformation, averaged them, and transformed them back to correlation coefficients.

RoBERTa fine tuning procedure. Robustly Optimized BERT Pre-training Approach (RoBERTa) (18) is an advanced language representation model considered a meaningful innovation upon prior algorithms in the field of natural language processing. It is a deep neural network that has been pre-trained by having it predict masked words in extremely large volumes of generic text (i.e., books and English Wikipedia). The fine-tuning process consists of adjusting the parameters of the final layers in order to maximize predictive accuracy in particular tasks (e.g., text classification), and in a particular corpus of text (e.g., admissions essays).

We used a subset of essays that were not manually coded to do a round of pre-training to optimize the RoBERTa model to our admission essay corpus. To do this, we trained RoBERTa to predict a masked word given the surrounding words. This process resulted in a RoBERTa model optimized for the particular prompt the essays in our corpus were answering. See **Supplementary Materials** for technical details on the pre-training process.

To begin the fine-tuning procedure, the second and third authors read random batches of 50 applicant essays to identify salient personal qualities commonly identified by colleges as desirable and/or shown in prior research to be related to positive life outcomes. After reading and discussing nine batches of 450 essays each, they developed criteria for seven personal qualities: prosocial purpose, leadership, teamwork, learning, perseverance, intrinsic motivation, and goal pursuit.

Next, we trained five research assistants to apply these criteria until each coder achieved adequate inter-rater reliability with either the second or third author across all seven attributes (Krippendorff's alpha $> .80$). Raters then coded all 3,131 essays in the *Development Sample.* Most of the essays were coded by a single rater ($n = 2{,}925$; 93% of the *Development Sample*). To assess inter-rater reliability, pairs of raters independently coded a subset of essays ($n = 206$; 7% of the *Development Sample*).

We used this manually annotated dataset to fine-tune separate RoBERTa models to estimate the probability of each personal quality. After fine-tuning these models, we evaluated the performance of the models and applied it to the hold-out sample of 306,463 essays, yielding more than two million continuous codes.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript. We thank Natalie Yee, Xena Wang, The Character Collaborative, NACAC, and Damien Crone for their help in this research

1. National Research Council, *Assessing 21st century skills: Summary of a workshop* (The National Academies Press, Washington DC, 2011). Publisher: National Academies Press.
2. T. E. Moffitt, L. Arseneault, D. Belsky, N. Dickson, R. J. Hancox, H. Harrington, R. Houts, R. Poulton, B. W. Roberts, S. Ross, M. R. Sears, W. M. Thomson, A. Caspi, A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* **108**, 2693–2698 (2011).
3. M. Almlund, A. L. Duckworth, J. Heckman, T. Kautz, Personality Psychology and Economics. *Handbook of the Economics of Education* (Elsevier, 2011), vol. 4, pp. 1–181.
4. S. B. Robbins, K. Lauver, H. Le, D. Davis, R. Langley, A. Carlstrom, Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychol. Bull.* **130**, 261–288 (2004).
5. A. L. Coleman, J. L. Keith, Understanding Holistic Review in Higher Education Admissions, *Tech. rep.*, College Board, New York (2018).
6. J. Karabel, *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton.* (Houghton Mifflin Harcourt, 2005).
7. T. R. Anderson, R. Weissbourd, Character Assessment in College Admission, *Tech. rep.*, Making Caring Common Project, Boston (2020).
8. D. Kahneman, O. Sibony, C. R. Sunstein, *Noise: A Flaw in Human Judgment* (Harper Collins, 2021).
9. J. P. Rushton, C. J. Brainerd, M. Pressley, Behavioral development and construct validity: The principle of aggregation. *Psychol. Bull.* **94**, 18 (1983).
10. E. Hoover, Working Smarter, Not Harder, in Admissions (2017). Section: News.
11. M. Korn, Some Elite Colleges Review an Application in 8 Minutes (or Less). *Wall Str. J.* (2018).
12. J. Manyika, J. Silberg, B. Presten, What Do We Do About the Biases in AI? p. 5 (2019).
13. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. p. 8 (2019).
14. D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, S. Venkatasubramanian, Runaway Feedback Loops in Predictive Policing. *Proc. Mach. Learn. Res.* **81**, 12 (2018).
15. A. Alvero, S. Giebel, B. Gebre-Medhin, a. l. antonio, M. L. Stevens, B. W. Domingue, Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Sci. Adv.* **7**, eabi9031 (2021).
16. M. O. Riedl, Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **1**, 33–36 (2019).
17. B. Shneiderman, Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interact.* pp. 109–124 (2020).
18. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). ArXiv:1907.11692 [cs].
19. C. Pierse, Transformers Interpret (2021). Original-date: 2020-05-27T20:32:08Z.
20. J. D. Janizek, P. Sturmfels, S.-I. Lee, Explaining Explanations: Axiomatic Feature Interactions for Deep Networks (2020). ArXiv:2002.04138 [cs, stat].
21. L. Kamas, A. Preston, Empathy, gender, and prosocial behavior. *J. Behav. Exp. Econ.* **92**, 101654 (2021).
22. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *The Annals Stat.* **29** (2001).
23. J. W. Pennebaker, C. K. Chung, J. Frazee, G. M. Lavergne, D. I. Beaver, When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLoS ONE* **9**, e115844 (2014).
24. W. W. Willingham, *Success in college: The role of personal qualities and academic ability* (College Board Publications, 1985).
25. M. J. Salganik, *et al.*, Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* **117**, 8398–8403 (2020).
26. T. Martin, J. M. Hofman, A. Sharma, A. Anderson, D. J. Watts, *Proceedings of the 25th International Conference on World Wide Web* (International World Wide Web Conferences Steering Committee, Montréal Québec Canada, 2016), pp. 683–694.
27. S. Ahadi, E. Diener, Multiple determinants and effect size. *J. Pers. Soc. Psychol.* **56**, 398–406 (1989).
28. S. Goldrick-Rab, Following Their Every Move: An Investigation of Social-Class Differences in College Pathways. *Sociol. Educ.* **79**, 67–79 (2006).
29. D. Hepworth, B. Littlepage, K. Hancock, Factors influencing university student academic success. *Educ. Res. Q.* **42**, 45–61 (2018).
30. S. F. Porchea, J. Allen, S. Robbins, R. P. Phelps, Predictors of Long-Term Enrollment and Degree Outcomes for Community College Students: Integrating Academic, Psychosocial, Socio-demographic, and Situational Factors. *The J. High. Educ.* **81**, 680–708 (2010).
31. M. C. Murphy, M. Gopalan, E. R. Carter, K. T. U. Emerson, B. L. Bottoms, G. M. Walton, A customized belonging intervention improves retention of socially disadvantaged students at a broad-access university. *Sci. Adv.* **6**, eaba4677 (2020).
32. R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, R. Tang, "Garbage In, Garbage Out" Revisited: What Do Machine Learning Application Papers Report About Human-Labeled Training Data? *Quant. Sci. Stud.* pp. 1–32 (2021). ArXiv: 2107.02278.
33. D. J. Benjamin, D. Laibson, W. Mischel, P. K. Peake, Y. Shoda, A. S. Wellsjo, N. L. Wilson, Predicting mid-life capital formation with pre-school delay of gratification and life-course measures of self-regulation. *J. Econ. Behav. & Organ.* **179**, 743–756 (2020).
34. A. L. Duckworth, M. E. Seligman, Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychol. Sci.* **16**, 939–944 (2005).
35. B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
36. D. T. Campbell, Assessing the impact of planned social change. *Eval. Program Plan.* **2**, 67–90 (1979).
37. S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** (2011).

Supplementary Materials for
Using Human-Centered Artificial Intelligence to Assess Personal Qualities in College Admissions

Authors

Benjamin Lira[1*], Margo Gardner[2], Abigail Quirk[1], Cathlyn Stone[2], Arjun Rao[2], Lyle Ungar[1], Stephen Hutt[1], Sidney K. D'Mello[2], and Angela L. Duckworth[1]

Affiliations

[1]University of Pennsylvania

[2]University of Colorado, Boulder

*Correspondence concerning this article should be addressed to Benjamin Lira, University of Pennsylvania.
Email: blira@sas.upenn.edu

Data and Exclusions

The dataset for this study emerged from a collaboration with the Common Application (Common App, www.commonapp.org) and the National Student Clearinghouse (NSC, www.studentclearinghouse.org). To protect privacy, Common App contracted a third-party organization to collect, anonymize, and deliver the dataset to our team. For additional details, see Hutt, et al. (*1*).

Specifically, our sample was drawn from the population of 413,675 students who completed the Common App during the 2008-09 academic year for college admission during the 2009-10 academic year. From this population, we selected the 311,308 students who had not enrolled in a postsecondary institution prior to 2008. This ensured the accuracy of records reflecting time to degree attainment.

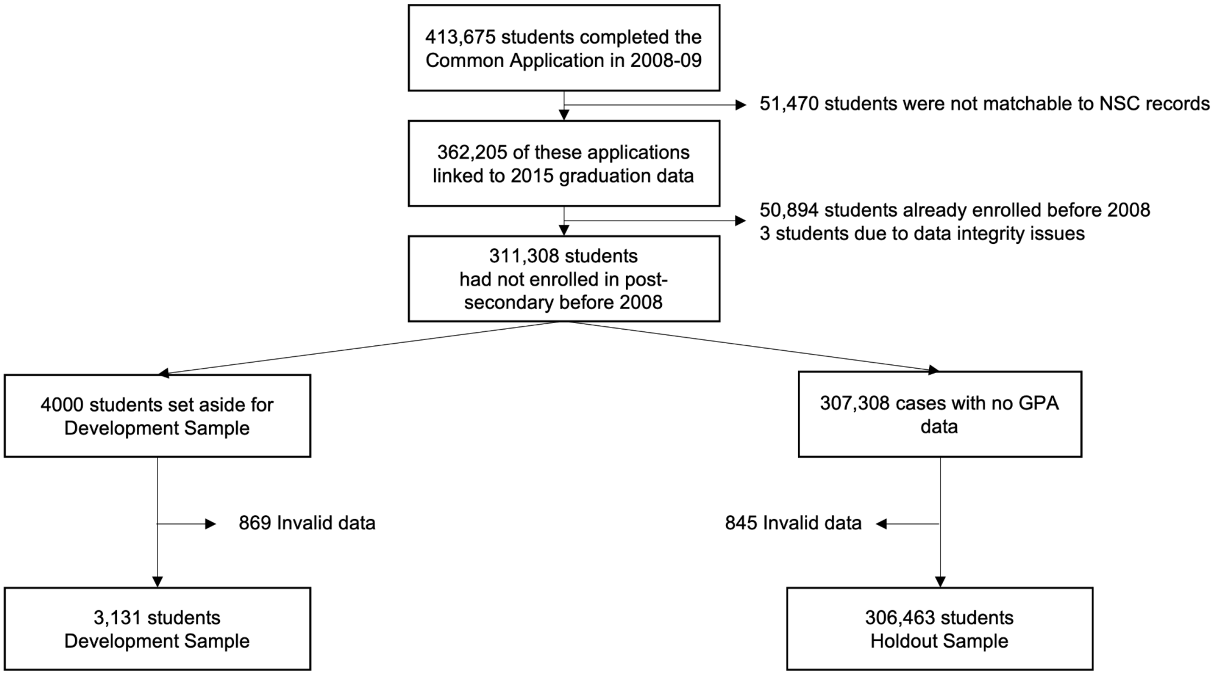
Development Sample

Originally, we identified a stratified sample of applications for manual coding. As reported previously (*1*), we defined sampling strata based on the number of extracurricular activities reported on the Common Application as well as membership in one of five multi-dimensional demographic groups identified using latent class analysis (LCA). Specifically, our LCA model classified students according to profiles across race/ethnicity, parental education, parents' marital status, English language learner status (ELL), attended a Title 1 high school, and high school race/ethnic composition. The LCA was performed in MPlus 7 on the subset of all 213,091 students attending public schools. We excluded private and homeschooled students from this analysis because their school-level demographic data were not available. After excluding missing data, invalid responses, and essays coded by one rater who ultimately failed to achieve agreement with other raters, the *Development Sample* consisted of 3,131 students.

Holdout Sample

Of the original 311,308 applications, we were then left with the remaining 307,308 applications, which were not manually coded. We excluded 54 cases for which the algorithm failed to generate computer likelihoods, suggesting data errors; 786 essays with fewer than 50 characters (most of which had no content, e.g.,"see attachment"); 3 applications with invalid essays (i.e., essays written by different applicants that were accidentally concatenated together); and 2 applications for which we had no available demographic information. This left us with a final *Holdout Sample* of 306,463 applicants. See **Figure S1** for a graphical representation of the sample composition.

Figure S1. Samples and exclusions


413,675 students completed the Common Application in 2008-09
51,470 students were not matchable to NSC records
362,205 of these applications linked to 2015 graduation data
50,894 students already enrolled before 2008
3 students due to data integrity issues
311,308 students had not enrolled in post-secondary before 2008
4000 students set aside for Development Sample
307,308 cases with no GPA data
869 Invalid data
845 Invalid data
3,131 students Development Sample
306,463 students Holdout Sample

RoBERTa Algorithm Fine-Tuning Procedure

We used the RoBERTa-base model, which we obtained from huggingface's "transformers" Python library. See this link to the model hosted on the huggingface website.

We began with pre-training, a procedure where the model is trained to identify words that have been removed from the text (i.e., masked language modeling). We used a single training epoch on unlabelled data to avoid overfitting.

We then finetuned the resulting model on our human-labelled dataset used 4 training epochs, with 32 examples used to predict on before updating the weights in each iteration (batch size = 32). See our settings in **Appendix A**.

We used a 10-fold cross-validation procedure for training the model. Specifically, the *Development Sample* of 3,131 hand-coded essays was divided into 10 random subsets. We fine-tuned RoBERTa models on nine subsets and generated predictions on the held-out subset. We repeated this process until each subset was used for testing once. We then pooled the computer-generated likelihoods over the 10 iterations. All measures of model accuracy are based on out-of-sample predictions.

We used a binary classification framework. Specifically, we separately fine-tuned 10 models (one for each subset of cross-validation) for each of the seven personal qualities. Our final RoBERTa procedure entails applying these 70 RoBERTa models to each application essay and pooling predictions from each of the models to generate seven computer-generated likelihoods of personal qualities, which we used in subsequent analyses.

Descriptive Statistics

Tables S1 and **S2** show descriptive statistics and correlations for the study variables in the *Development* and *Holdout Samples.*

Table S1. Correlations and descriptive statistics in the *Development Sample*

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Prosocial purpose												
2. Leadership	-.02											
3. Teamwork	-.08***	.19***										
4. Learning	-.11***	-.02	.06***									
5. Perseverance	-.17***	.00	.05**	.13***								
6. Intrinsic motivation	-.06**	-.11***	.00	-.03	.06**							
7. Goal pursuit	.05**	.07***	.07***	-.05*	.05**	.04*						
8. Standardized test scores	-.01	.08***	.05**	.02	.08***	.00	.03					
9. Number of activities	.09***	.13***	.11***	.06***	.07***	-.01	.08***	.36***				
10. Time per activity	-.01	.13***	.05**	.01	.06**	.05**	.07***	.26***	.40***			
11. Proportion sports	-.12***	.03	.04*	.01	.05**	.08***	.03	-.11***	-.06**	.28***		
12. College graduation	.03	.07***	.05**	.06***	.03	.02	.03	.30***	.22***	.17***	.00	
M	0.36	0.19	0.26	0.45	0.19	0.45	0.32	1,693	3.46	2.11	0.23	0.66
SD	0.46	0.37	0.39	0.47	0.35	0.45	0.40	306	2.29	1.13		
N	3,120	3,124	3,103	3,126	3,125	3,116	3,124	2,834	3,131	3,131	3,131	3,131

Note. Correlations with computer-generated scores of personal qualities below the diagonal, correlations with human ratings of personal qualities above the diagonal.
*** *p* < .001. ** *p* < .01. * *p* < .05.

Table S2. Correlations and descriptive statistics in the *Holdout Sample*

Variable	1	2	3	4	5	6	7	8	9	10	11	12
1. Prosocial purpose												
2. Leadership	-.02***											
3. Teamwork	-.14***	.19***										
4. Learning	-.13***	-.03***	.06***									
5. Perseverance	-.18***	-.02***	.07***	.12***								
6. Intrinsic motivation	-.08***	-.12***	-.01***	-.06***	.04***							
7. Goal pursuit	.08***	.07***	-.01***	-.08***	.01***	.02***						
8. Standardized test scores	.00*	.07***	.06***	.00*	.07***	.02***	.05***					
9. Number of activities	.10***	.09***	.06***	.03***	.05***	.03***	.08***	.34***				
10. Time per activity	-.03***	.07***	.02***	-.01**	.02***	.06***	.00	.14***	.03***			
11. Proportion sports	-.10***	-.03***	.02***	.01***	.05***	.03***	-.03***	-.21***	-.27***	.09***		
12. College graduation	.04***	.05***	.04***	.02***	.02***	.02***	.02***	.22***	.18***	.09***	-.05***	
M	0.37	0.20	0.30	0.47	0.21	0.51	0.36	1,826	5.16	2.53	0.26	0.78
SD	0.46	0.37	0.39	0.46	0.35	0.44	0.40	267	1.98	0.76		

Note. *** *p* < .001. ** *p* < .01. * *p* < .05. *N* = 306,463 for all variables other than standardized test scores (*n* = 289,140)

Relationship Between Personal Qualities And Demographics

As shown in **Table S3,** demographic subgroup differences in the binary human ratings of personal qualities were small in magnitude (and in most cases not reliably different from zero) in the *Development Sample.* As shown in **Table S4** and **S5**, these differences were likewise small for the continuous computer-generated likelihoods of personal qualities in the *Development Sample* and *Holdout Sample*.

Table S3. Human ratings of personal qualities by demographic subgroup in the *Development Sample*

Demographic variable	PP	LD	TW	LR	PS	IM	GP
Race/ethnicity							
White	−0.05	0.04	0.00	0.01	0.00	0.03	0.03
Black	−0.01	−0.02	−0.03	−0.03	−0.05	−0.01	0.00
Latino	0.03	0.00	0.01	−0.01	0.03	0.00	−0.03
Asian	0.05	−0.02	0.05	0.05	0.04	−0.04	0.00
Other	0.01	0.02	−0.04	−0.05	−0.04	0.01	0.02
Missing	−0.03	−0.02	0.01	0.01	0.01	0.01	−0.04
Number of parents with college degrees							
None	0.03	−0.01	−0.03	−0.04	−0.04	−0.01	−0.04
One	−0.03	−0.01	0.00	0.00	−0.01	0.00	0.01
Two	−0.01	0.03	0.03	0.04	0.05	0.01	0.03
Female	0.13	0.03	0.04	0.04	0.04	0.05	−0.01
Married parents	0.00	0.02	0.00	0.00	0.02	0.02	0.02
English language learner	0.05	−0.03	0.03	0.02	0.03	−0.04	0.01
Title 1 High School	0.01	0.03	−0.01	−0.02	−0.01	0.01	−0.01

Note. PP, Prosocial purpose; LD, Leadership; TW, Teamwork; LR, Learning; PS, Perseverance; IM, Intrinsic motivation; GP, Goal pursuit. Values are Matthew's correlation coefficients (phi)

Table S4. Computer-generated likelihoods of personal qualities by demographic subgroup in the *Development Sample*

Demographic variable	PP	LD	TW	LR	PS	IM	GP
Race/ethnicity							
White	−0.10	0.11	0.03	−0.02	0.02	0.15	0.05
Black	0.01	−0.06	−0.10	−0.07	−0.14	−0.04	0.01
Latino	0.09	−0.04	0.04	−0.02	0.04	−0.02	−0.01
Asian	0.14	−0.04	0.06	0.14	0.10	−0.16	−0.02
Other	0.02	0.03	−0.11	−0.16	−0.15	−0.04	0.05
Missing	−0.15	−0.07	0.02	0.08	0.07	0.05	−0.11
Number of parents with college degrees							
None	0.09	−0.03	−0.15	−0.07	−0.05	−0.01	−0.04
One	−0.06	0.01	0.07	−0.01	−0.04	0.00	−0.04
Two	−0.06	0.04	0.13	0.10	0.10	0.01	0.09
Female	0.26	0.06	0.07	0.08	0.09	0.14	0.06
Married parents	−0.03	0.04	0.04	0.00	0.07	0.02	0.07
English language learner	0.13	−0.07	0.03	0.05	0.08	−0.10	0.06
Title 1 High School	0.03	0.02	0.02	−0.04	−0.03	0.03	−0.04

Note. PP, Prosocial purpose; LD, Leadership; TW, Teamwork; LR, Learning; PS, Perseverance; IM, Intrinsic motivation; GP, Goal pursuit. Values are Cohen's *d*s

Table S5. Computer-generated likelihoods of personal qualities by demographic subgroup in the *Holdout Sample*

Demographic variable	PP	LD	TW	LR	PS	IM	GP
Race/ethnicity							
White	−0.03	0.04	0.04	0.01	−0.01	0.07	−0.02
Black	0.01	−0.04	−0.11	−0.10	−0.13	−0.14	−0.08
Latino	0.08	−0.04	−0.09	−0.01	−0.07	−0.07	−0.04
Asian	0.07	0.00	−0.01	0.08	0.10	−0.12	0.08
Other	0.00	−0.01	−0.04	−0.02	−0.02	0.00	0.02
Missing	−0.03	−0.03	0.02	−0.02	0.04	0.03	0.01
Number of parents with college degrees							
None	0.00	−0.06	−0.09	−0.04	−0.10	−0.08	−0.08
One	−0.01	−0.01	−0.01	0.00	−0.03	0.00	0.00
Two	0.01	0.06	0.08	0.03	0.10	0.06	0.07
Female	0.22	0.06	0.01	0.03	0.02	0.14	0.03
Married parents	0.05	0.06	0.05	0.03	0.06	0.02	0.05
English language learner	0.06	−0.06	−0.05	0.05	0.06	−0.11	0.06
Title 1 high school	−0.02	0.02	−0.01	0.00	−0.03	−0.06	−0.02

Note. PP, Prosocial purpose; LD, Leadership; TW, Teamwork; LR, Learning; PS, Perseverance; IM, Intrinsic motivation; GP, Goal pursuit. Values are Cohen's *d*s

Human-Computer Correlations Across Demographic Subgroups

As shown in **Table S6,** the convergent validity for each group was, for the most part, not significantly different from the convergent validity of the most populated subgroup. **Table S6** shows the correlation between human ratings and computer-generated likelihoods of personal qualities for each subgroup compared to the reference group.

Table S6. Difference between human-computer correlations for each subgroup compared to the reference group.

	PP			LD			TW			LR			PS			IM			GP		
	C	R	D	C	R	D	C	R	D	C	R	D	C	R	D	C	R	D	C	R	D
Race/ethnicity (vs. White)																					
Black	.84	.87	-.04	.78	.79	-.01	.64	.59	.05	.76	.80	-.04	.72	.68	.04	.78	.73	.05	.60	.58	.01
Latino	.86	.87	-.02	.85	.79	.06*	.60	.59	.02	.78	.80	-.03	.63	.68	-.05	.71	.73	-.02	.67	.58	.08
Asian	.84	.87	-.03	.80	.79	.01	.62	.59	.03	.72	.80	-.08**	.67	.68	-.01	.70	.73	-.02	.54	.58	-.04
Other	.84	.87	-.03	.83	.79	.04	.61	.59	.02	.72	.80	-.08*	.56	.68	-.12*	.74	.73	.01	.54	.58	-.05
No race reported	.90	.87	.02	.85	.79	.06*	.66	.59	.07	.77	.80	-.04	.70	.68	.02	.75	.73	.02	.65	.58	.06
Number of parents with college degrees (vs. None)																					
One	.86	.85	.01	.81	.81	-.00	.62	.60	.02	.79	.78	.01	.65	.66	-.01	.71	.75	-.04	.57	.61	-.04
Two	.88	.85	.03*	.80	.81	-.01	.63	.60	.02	.74	.78	-.04	.69	.66	.03	.71	.75	-.04	.58	.61	-.04
Other demographics																					
Male	.85	.86	-.01	.81	.81	.00	.61	.62	-.01	.77	.77	-.00	.66	.68	-.02	.74	.73	.01	.58	.60	-.02
Parents not married	.88	.85	.03*	.80	.81	-.02	.63	.61	.02	.77	.77	.01	.66	.68	-.02	.71	.74	-.02	.62	.58	.04
English language learner	.87	.86	.02	.77	.82	-.05*	.59	.62	-.03	.74	.78	-.03	.69	.66	.02	.72	.73	-.02	.61	.59	.02
Title 1 public school	.88	.83	.04**	.80	.84	-.04*	.63	.61	.02	.78	.77	.01	.66	.67	-.01	.72	.74	-.02	.60	.59	.01

Note. PP, Prosocial purpose; LD, Leadership; TW, Teamwork; LR, Learning; PS, Perseverance; IM, Intrinsic motivation; GP, Goal pursuit. C, Human-computer correlation for the comparison group; R, Human-computer correlation for the reference group; D, Differences between point-biserial correlations. *p*-values are adjusted for multiple comparisons using the Benjamini Hochberg False Discovery Rate correction (*2*).

Tables S7 to **S23** show descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities for each of 17 subgroups defined by personal characteristics (i.e., gender, parental education, parental marital status, English language learner status, race/ethnicity, and type of high school).

Table S7. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for White applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.87***	-.03	-.07*	-.08*	-.11***	-.04	.07*
2. Leadership	-.04	.79***	.08*	-.04	-.03	-.13***	.01
3. Teamwork	-.10**	.11***	.59***	.07*	.04	-.01	.03
4. Learning	-.15***	-.04	.04	.80***	.14***	-.04	-.07*
5. Perseverance	-.15***	-.07*	.05	.09**	.68***	.04	.06
6. Intrinsic motivation	-.04	-.13***	.04	-.05	.04	.73***	.02
7. Goal pursuit	.10**	-.01	.02	.01	.00	.04	.58***
Frequency of human rating	0.31	0.20	0.26	0.42	0.19	0.44	0.34
Mean of computer-generated likelihood	0.33	0.22	0.27	0.44	0.19	0.50	0.34

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S8. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for Black applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.84***	-.04	.03	-.06	-.22***	-.04	-.03
2. Leadership	.04	.78***	.18***	-.07	-.08	-.06	.06
3. Teamwork	.05	.22***	.64***	.02	.02	-.02	.07
4. Learning	-.06	-.04	.03	.76***	.04	-.02	-.03
5. Perseverance	-.27***	-.09*	.06	.09	.72***	.01	-.02
6. Intrinsic motivation	-.09*	-.02	.01	-.07	.05	.78***	.10*
7. Goal pursuit	.00	.03	.15**	-.03	-.02	.02	.60***
Frequency of human rating	0.34	0.16	0.22	0.38	0.14	0.40	0.32
Mean of computer-generated likelihood	0.36	0.18	0.23	0.42	0.15	0.44	0.33

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S9. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for Latino applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.86***	.03	-.02	-.14**	-.13**	-.04	.06
2. Leadership	.00	.85***	.19***	-.02	.10*	.01	.10*
3. Teamwork	-.09*	.21***	.60***	.08	.17***	.00	.10*
4. Learning	-.12**	-.07	.09*	.78***	.10*	.00	-.01
5. Perseverance	-.12**	.10*	.08	.10*	.63***	.05	.08
6. Intrinsic motivation	-.02	-.05	-.06	.01	.02	.71***	-.01
7. Goal pursuit	.06	.10*	.08	-.08	.09*	.01	.67***
Frequency of human rating	0.37	0.18	0.27	0.41	0.22	0.41	0.28
Mean of computer-generated likelihood	0.39	0.18	0.27	0.44	0.20	0.44	0.32

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S10. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for Asian applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.84***	.03	-.11**	-.08*	-.07	-.04	.04
2. Leadership	-.02	.80***	.20***	.01	.02	-.15***	.05
3. Teamwork	-.15***	.18***	.62***	.06	.06	-.01	.07
4. Learning	-.08*	-.12**	.00	.72***	.08*	-.02	.00
5. Perseverance	-.14***	.04	.06	.09*	.67***	.11**	.11**
6. Intrinsic motivation	-.05	-.10*	-.01	-.05	.14***	.70***	-.02
7. Goal pursuit	.04	.16***	.10*	-.01	.05	.01	.54***
Frequency of human rating	0.40	0.16	0.30	0.47	0.22	0.38	0.32
Mean of computer-generated likelihood	0.41	0.18	0.28	0.50	0.21	0.39	0.32

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

Table S11. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants reporting other races/ethnicities

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.84***	.03	-.04	-.16**	-.21***	-.05	-.01
2. Leadership	-.01	.83***	.14*	.05	.01	-.17**	.08
3. Teamwork	-.05	.23***	.61***	.01	.07	-.08	.08
4. Learning	-.07	-.03	.05	.72***	.08	.08	.07
5. Perseverance	-.15**	-.07	.06	.14*	.56***	-.08	.06
6. Intrinsic motivation	.06	-.15*	-.07	.09	-.10	.74***	.10
7. Goal pursuit	.06	.08	.04	.09	-.06	.03	.54***
Frequency of human rating	0.36	0.20	0.20	0.35	0.14	0.43	0.34
Mean of computer-generated likelihood	0.37	0.20	0.22	0.38	0.14	0.44	0.34

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

Table S12. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants who did not report their race/ethnicity

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.90***	-.02	.02	-.07	-.09	-.10*	.11*
2. Leadership	.01	.85***	.20***	.06	.01	-.04	.02
3. Teamwork	.00	.20***	.66***	.05	.01	.00	.01
4. Learning	-.08	.02	-.02	.77***	.15**	.00	-.05
5. Perseverance	-.14**	.03	.00	.12*	.70***	-.02	.01
6. Intrinsic motivation	-.09	-.05	.02	-.07	-.02	.75***	.04
7. Goal pursuit	.11*	.04	-.02	-.03	.07	.01	.65***
Frequency of human rating	0.30	0.15	0.26	0.43	0.20	0.43	0.26
Mean of computer-generated likelihood	0.30	0.17	0.27	0.49	0.21	0.47	0.28

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

Table S13. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants with no parents with college degrees

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.85***	-.03	-.03	-.07**	-.13***	-.04	.05*
2. Leadership	-.02	.81***	.17***	.02	.02	-.07**	.06*
3. Teamwork	-.09***	.20***	.60***	.07**	.05	-.02	.03
4. Learning	-.09***	-.03	.07**	.78***	.08**	-.02	-.02
5. Perseverance	-.19***	.02	.05	.07**	.66***	.02	.03
6. Intrinsic motivation	-.03	-.07**	-.04	-.04	.02	.75***	.04
7. Goal pursuit	.05*	.07**	.06*	-.03	.03	.04	.61***
Frequency of human rating	0.36	0.17	0.24	0.40	0.17	0.42	0.30
Mean of computer-generated likelihood	0.38	0.19	0.23	0.44	0.18	0.45	0.32

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S14. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants with one parent with a college degree

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.86***	.05	-.07	-.09*	-.09*	-.03	.03
2. Leadership	.02	.81***	.07	-.05	-.04	-.11**	.05
3. Teamwork	-.05	.08*	.62***	.06	.09*	.04	.10*
4. Learning	-.06	-.07	.04	.79***	.13***	.01	-.05
5. Perseverance	-.12**	-.08*	.07	.17***	.65***	.01	.03
6. Intrinsic motivation	-.04	-.11**	.11**	-.04	.08*	.71***	.04
7. Goal pursuit	.07	.02	.03	.00	-.04	.00	.57***
Frequency of human rating	0.32	0.17	0.26	0.41	0.18	0.41	0.32
Mean of computer-generated likelihood	0.34	0.19	0.28	0.45	0.18	0.45	0.31

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S15. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants with two parents with college degrees

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.88***	-.01	-.04	-.12***	-.13***	-.09**	.04
2. Leadership	-.02	.80***	.18***	-.03	.00	-.11**	.02
3. Teamwork	-.03	.19***	.63***	.02	.06	-.06	.06
4. Learning	-.14***	-.07*	-.03	.74***	.15***	-.03	-.04
5. Perseverance	-.13***	-.02	.07	.11**	.69***	.07*	.10**
6. Intrinsic motivation	-.09**	-.10**	-.03	-.02	.05	.71***	.00
7. Goal pursuit	.08*	.07*	.08*	.01	.05	.01	.58***
Frequency of human rating	0.33	0.19	0.28	0.45	0.22	0.43	0.34
Mean of computer-generated likelihood	0.34	0.20	0.30	0.48	0.21	0.46	0.35

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S16. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for female applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.86***	.00	-.05*	-.10***	-.16***	-.02	.04
2. Leadership	.00	.81***	.16***	.00	.02	-.09***	.03
3. Teamwork	-.10***	.18***	.62***	.05*	.07**	-.03	.04
4. Learning	-.10***	-.04	.06*	.77***	.11***	-.02	-.02
5. Perseverance	-.18***	.00	.07**	.09***	.68***	.03	.05*
6. Intrinsic motivation	-.03	-.11***	-.01	-.05	.02	.73***	.02
7. Goal pursuit	.08**	.06*	.08***	-.01	.02	.03	.60***
Frequency of human rating	0.40	0.19	0.27	0.44	0.20	0.44	0.31
Mean of computer-generated likelihood	0.41	0.20	0.27	0.47	0.20	0.48	0.33

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S17. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for male applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.85***	-.02	-.04	-.10***	-.09***	-.11***	.04
2. Leadership	-.03	.81***	.14***	-.03	-.02	-.09***	.07**
3. Teamwork	-.05	.17***	.61***	.07*	.06*	-.01	.08**
4. Learning	-.12***	-.07*	.00	.77***	.11***	-.01	-.04
5. Perseverance	-.15***	-.04	.04	.12***	.66***	.03	.05
6. Intrinsic motivation	-.10***	-.06*	.00	-.03	.06*	.74***	.05
7. Goal pursuit	.03	.05	.03	-.01	.02	.01	.58***
Frequency of human rating	0.27	0.17	0.24	0.39	0.17	0.39	0.32
Mean of computer-generated likelihood	0.29	0.18	0.25	0.43	0.17	0.42	0.31

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S18. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants with married parents

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.85***	-.01	-.04*	-.12***	-.12***	-.04*	.04
2. Leadership	-.01	.81***	.15***	-.01	.02	-.09***	.05*
3. Teamwork	-.07**	.17***	.61***	.06**	.08***	.00	.05*
4. Learning	-.12***	-.06**	.02	.77***	.12***	-.03	-.03
5. Perseverance	-.15***	-.01	.05*	.10***	.68***	.03	.06**
6. Intrinsic motivation	-.06**	-.09***	.02	-.05*	.03	.74***	.02
7. Goal pursuit	.05*	.07**	.05*	-.02	.03	.01	.58***
Frequency of human rating	0.34	0.18	0.26	0.42	0.19	0.42	0.32
Mean of computer-generated likelihood	0.35	0.20	0.27	0.45	0.20	0.46	0.33

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S19. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants with parents who are not married

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.88***	.00	-.04	-.04	-.13***	-.06*	.05
2. Leadership	-.01	.80***	.15***	-.02	-.02	-.08**	.05
3. Teamwork	-.07*	.19***	.63***	.05	.04	-.04	.06*
4. Learning	-.07*	-.03	.07*	.77***	.08**	.01	-.02
5. Perseverance	-.18***	-.01	.07*	.11***	.66***	.05	.04
6. Intrinsic motivation	-.03	-.07*	-.04	-.01	.06	.71***	.06
7. Goal pursuit	.08**	.04	.08**	.01	.02	.04	.62***
Frequency of human rating	0.35	0.17	0.25	0.42	0.18	0.41	0.30
Mean of computer-generated likelihood	0.37	0.18	0.25	0.45	0.17	0.45	0.31

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

Table S20. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for English language learner applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.87***	.05	-.05	-.14***	-.09**	-.03	.02
2. Leadership	.03	.77***	.13***	.00	.00	-.08*	.06
3. Teamwork	-.08*	.17***	.59***	.06	.10**	-.01	.07*
4. Learning	-.15***	-.09**	.06	.74***	.08*	.01	-.02
5. Perseverance	-.15***	.02	.05	.10**	.69***	.03	.09*
6. Intrinsic motivation	-.03	-.09*	-.01	-.01	.03	.72***	-.01
7. Goal pursuit	.02	.11**	.08*	-.02	.04	.01	.61***
Frequency of human rating	0.38	0.16	0.28	0.43	0.21	0.39	0.32
Mean of computer-generated likelihood	0.40	0.17	0.27	0.47	0.21	0.42	0.34

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

Table S21. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for native English-speaking applicants

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.86***	-.02	-.04	-.08***	-.14***	-.06**	.05*
2. Leadership	-.03	.82***	.16***	-.02	.00	-.10***	.05*
3. Teamwork	-.07**	.18***	.62***	.06**	.05*	-.02	.05*
4. Learning	-.08***	-.04	.03	.78***	.12***	-.02	-.03
5. Perseverance	-.17***	-.02	.06**	.10***	.66***	.04	.04
6. Intrinsic motivation	-.05**	-.09***	.00	-.04*	.04*	.73***	.05*
7. Goal pursuit	.07***	.04*	.05*	-.01	.02	.03	.59***
Frequency of human rating	0.33	0.18	0.25	0.41	0.18	0.43	0.31
Mean of computer-generated likelihood	0.34	0.20	0.26	0.45	0.18	0.46	0.32

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

Table S22. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants who attended Title 1 public high schools

Personal quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.83***	.00	-.04	-.05	-.18***	-.05	.05
2. Leadership	.02	.84***	.19***	-.03	.05	-.07*	.06*
3. Teamwork	-.08*	.21***	.61***	.04	.12***	-.04	.07*
4. Learning	-.08**	-.06*	.03	.77***	.10***	-.01	-.06
5. Perseverance	-.21***	.02	.08**	.11***	.67***	.04	.06*
6. Intrinsic motivation	-.07*	-.11***	-.04	-.06	.03	.74***	.04
7. Goal pursuit	.05	.08**	.07*	-.04	.06*	.05	.59***
Frequency of human rating	0.35	0.20	0.25	0.41	0.18	0.43	0.30
Mean of computer-generated likelihood	0.36	0.21	0.27	0.44	0.19	0.47	0.31

Note. *** p < .001. ** p < .01. * p < .05.

Table S23. Descriptive statistics and point biserial correlations between computer-generated likelihoods and human ratings of personal qualities in the *Development Sample* for applicants who attended non-Title-1 high schools

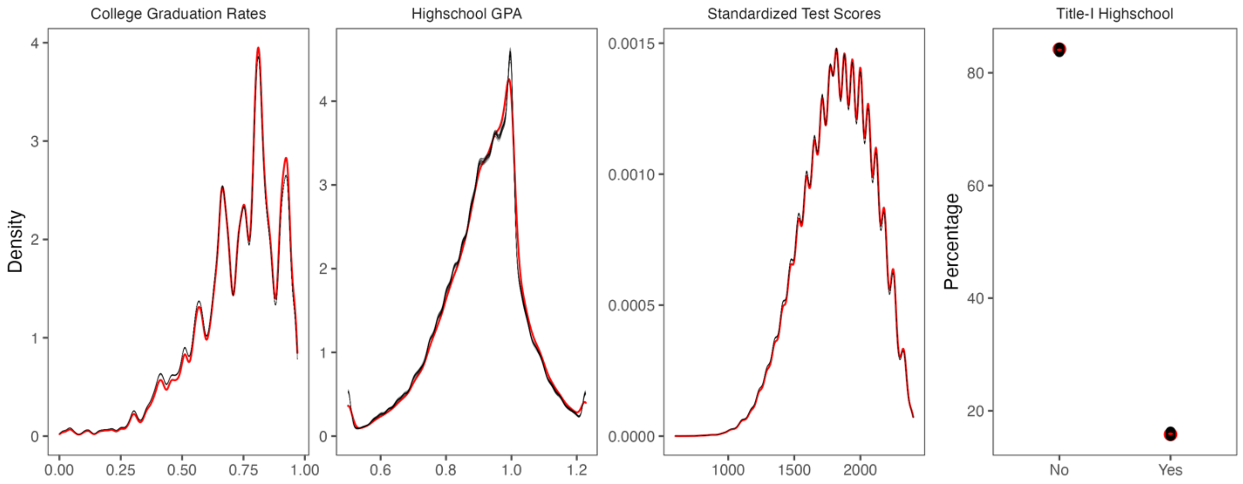
Personal Quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.88***	-.02	-.06*	-.12***	-.09***	-.05	.03
2. Leadership	-.03	.80***	.14***	.00	-.03	-.12***	.04
3. Teamwork	-.07**	.16***	.63***	.06*	.02	-.01	.06*
4. Learning	-.11***	-.03	.04	.78***	.13***	-.03	-.02
5. Perseverance	-.12***	-.02	.05*	.09***	.66***	.02	.04
6. Intrinsic motivation	-.03	-.08**	.02	-.02	.04	.72***	.03
7. Goal pursuit	.05	.05	.05*	.02	-.01	.01	.60***
Frequency of human rating	0.34	0.18	0.26	0.43	0.19	0.42	0.31
Mean of computer-generated likelihood	0.35	0.20	0.26	0.46	0.20	0.45	0.32

Note. *** p < .001. ** p < .01. * p < .05.

Quality Check of Imputation Procedure For Missing Data

Figure S3 shows distributions of each of the variables with missing data. We show the original distributions in black, and the overlapping red distributions represent the m = 25 imputed datasets. As shown in **Figure S3,** imputed distributions closely resemble the original distribution, suggesting adequate imputation quality.

Figure S3. Imputation quality



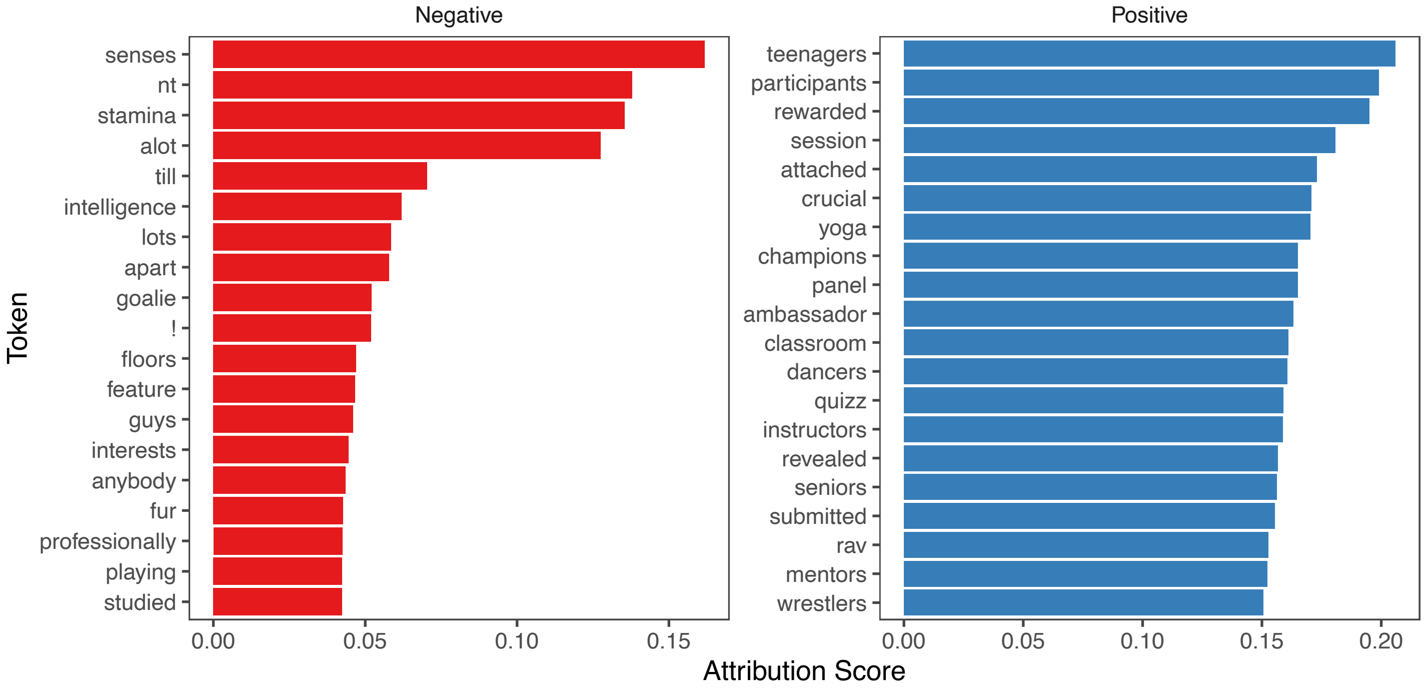
Details on Models Predicting Graduation Directly From Text

To estimate a ceiling on how much language could be predictive of graduation, we fine-tuned a model where student writing was used to predict whether students graduated. We used a random subset of 90% of the Holdout Sample for training, and the remaining 10% for testing the out-of-sample AUC of the model (AUC = .626).

Figure S4 shows the word tokens with the highest positive and negative attribution scores, that is the words and fractions of words that the model tended to use to classify students. Interestingly, misspellings (e.g., "alot"), exclamation marks ("!"), and informal language (e.g., "guys") tend to receive negative attribution scores.

Figure S4

Attribution scores for word tokens in a model where student writing directly predicts graduation.

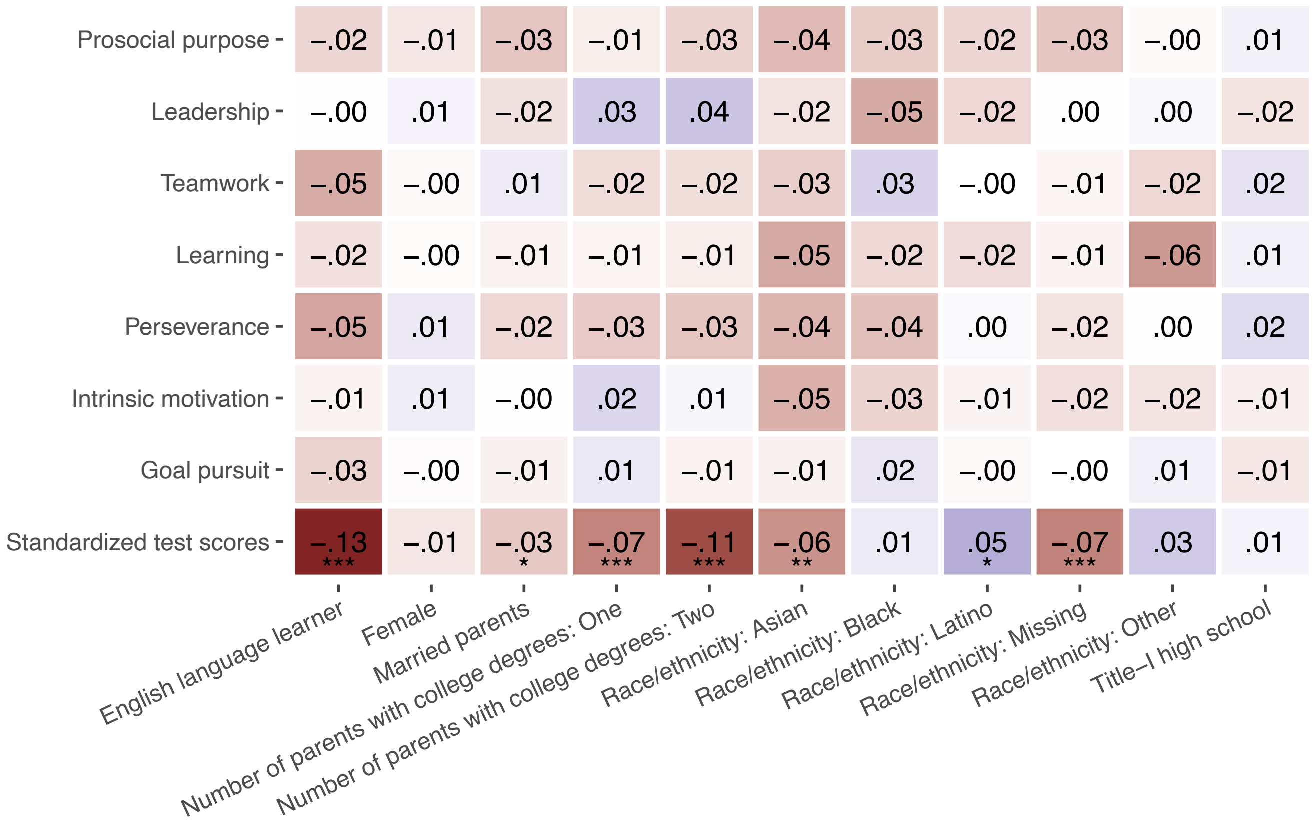


Details on Interaction Effects Of Demographics With Personal Qualities Predicting Graduation

We tested the equality of predictive validity of computer-generated likelihoods of personal qualities by fitting model (2) in **Table 4** in the main text but including interaction terms between each personal quality and standardized test scores and each demographic characteristic. **Figure S5** below shows the coefficients for each interaction term with statistical significance denoted with asterisks. These coefficients should be interpreted as the difference between the coefficient for the reference class, and the specified demographic category. For example, the coefficient in the top left indicates that prosocial purpose is .02 less predictive for English language learners as opposed to native speakers, the lack of asterisks means that the difference is not significant.

Figure S5

Coefficients of interaction terms between personal qualities and demographic characteristics in the prediction of six-year college graduation.



Robustness Checks For Analyses Predicting College Graduation

Predictive Validity Of Human Ratings Of Personal Qualities In The Development Sample

As shown in **Table S24,** in binary logistic regression models predicting college graduation, coefficients for human ratings of personal qualities in the *Development Sample* were similar to those of computer-generated likelihoods in the *Holdout Sample.*

Predictive Validity Of Computer-Generated Likelihoods Of Personal Qualities Controlling For High School GPA In The Holdout Sample

In the year of data collection, high school counselors had the option to submit report card grades either online or by uploading hard-copy transcripts. Because hard-copy transcripts were not possible to de-identify, we had access to only a subset of n = 43,597 applications in the holdout sample with high school grade point average (HSGPA). **Table S25** shows results of our main model specification including HSGPA as a predictor in that subsample.

Predictive Validity Of Computer-Generated Likelihoods Of Personal Qualities Controlling For Institutional Graduation Rates In The Holdout Sample

As shown in **Table S26,** in binary logistic regression models predicting college graduation, coefficients for computer-generated likelihoods of personal qualities were similar in magnitude to those presented in the main text.

Table S24. Binary logistic regression models predicting college graduation from human ratings of personal qualities in the *Development Sample*

	(1)	(2)
Human ratings of personal qualities		
Prosocial purpose	1.063	1.059
	(0.041)	(0.052)
Leadership	1.174***	1.066
	(0.048)	(0.052)
Teamwork	1.011	0.954
	(0.040)	(0.045)
Mastery orientation	1.137***	1.126*
	(0.044)	(0.054)
Perseverance	1.048	0.988
	(0.041)	(0.047)
Intrinsic motivation	1.037	1.055
	(0.040)	(0.050)
Goal pursuit	1.051	1.011
	(0.041)	(0.048)
Race/ethnicity (vs. White)		
Black		1.150
		(0.180)
Latino		0.776
		(0.126)
Asian		1.000
		(0.173)
Other		0.954
		(0.166)
No race reported		0.789
		(0.127)
Parental education (vs. no parent w/ college degree)		
One parent w/ college degree		1.282
		(0.163)
Two parents w/ college degree		1.544**
		(0.210)
Female		1.482***
		(0.148)
Married parents		1.138
		(0.117)
English language learner		1.164
		(0.160)
Title 1 high school		1.181
		(0.121)
Out-of-school activities (OSA)		
Number of OSA		1.199**
		(0.066)
Time per OSA		1.079
		(0.058)
Proportion sports		1.111*
		(0.056)
Standardized test scores		1.866***
		(0.116)
Constant	1.975***	1.459**
	(0.076)	(0.208)
N	3,078	2,484
AUC	.565	.719

Note. *** *p* < .001. ** *p* < .01. * *p* < .05.

Table S25. Binary logistic regression models predicting college graduation from computer-generated likelihoods of personal qualities controlling for high school GPA in the *Holdout Sample*

	(1)	(6)
Human ratings of personal qualities		
Prosocial purpose	1.113***	1.068***
	(0.081)	(0.090)
Leadership	1.132***	1.063***
	(0.081)	(0.090)
Teamwork	1.040**	0.993
	(0.079)	(0.089)
Mastery orientation	1.054***	1.038**
	(0.078)	(0.085)
Perseverance	1.072***	1.019
	(0.085)	(0.088)
Intrinsic motivation	1.063***	1.011
	(0.077)	(0.085)
Goal pursuit	1.046***	1.019
	(0.078)	(0.087)
Race/ethnicity (vs. White)		
Black		0.819***
		(0.370)
Latino		0.935
		(0.343)
Asian		0.760***
		(0.319)
Other		0.758***
		(0.299)
No race reported		0.839***
		(0.233)
Parental education (vs. no parent w/ college degree)		
One parent w/ college degree		1.261***
		(0.224)
Two parents w/ college degree		1.455***
		(0.220)
Female		1.385***
		(0.180)
Married parents		1.339***
		(0.200)
English language learner		0.711***
		(0.291)
Title 1 high school		0.909**
		(0.218)
Out-of-school activities		
Number of OSA		1.203***
		(0.088)
Time per OSA		1.081***
		(0.081)
Proportion sports		1.063***
		(0.085)
Standardized test scores		1.242***
		(0.104)
HSGPA		1.441***
		(0.096)
Constant	3.480***	2.446***
	(0.077)	(0.266)
AUC	.554	.702
N	43,591	43,591

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

Table S26. Binary logistic regression models predicting college graduation controlling for institutional graduation rates from human ratings of personal qualities in the *Holdout Sample*

	(1)	(6)
Human ratings of personal qualities		
Prosocial purpose	1.132***	1.063***
	(0.080)	(0.090)
Leadership	1.133***	1.055***
	(0.082)	(0.096)
Teamwork	1.080***	1.023***
	(0.083)	(0.093)
Mastery orientation	1.065***	1.036***
	(0.079)	(0.087)
Perseverance	1.071***	1.000
	(0.082)	(0.089)
Intrinsic motivation	1.068***	1.005
	(0.078)	(0.087)
Goal pursuit	1.041***	0.995
	(0.080)	(0.085)
Race/ethnicity (vs. White)		
Black		0.754***
		(0.328)
Latino		0.857***
		(0.306)
Asian		0.696***
		(0.350)
Other		0.733***
		(0.305)
No race reported		0.828***
		(0.244)
Parental education (vs. no parent w/ college degree)		
One parent w/ college degree		1.156***
		(0.231)
Two parents w/ college degree		1.196***
		(0.234)
Female		1.465***
		(0.183)
Married parents		1.281***
		(0.196)
English language learner		0.684***
		(0.290)
Title 1 high school		0.974*
		(0.230)
Institutional graduation rates		1.891***
		(0.095)
Out-of-school activities		
Number of OSA		1.159***
		(0.089)
Time per OSA		1.082***
		(0.082)
Proportion sports		1.029***
		(0.084)
Standardized test scores		1.164***
		(0.107)
Constant	3.558***	2.986***
	(0.004)	(0.277)
AUC	.560	.741
N	306,463	306,463

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

References

1. S. Hutt, M. Gardener, D. Kamentz, A. L. Duckworth, S. K. D'Mello, "Prospectively predicting 4-year college graduation from student applications" in *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (ACM, Sydney New South Wales Australia, 2018; https://dl.acm.org/doi/10.1145/3170358.3170395), pp. 280–289.
2. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29** (2001), doi:10.1214/aos/1013699998.