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

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

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

This manuscript was compiled on August 20, 2022

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 resource-intensive and idiosyncratic. On the other hand, unsupervised artificial intelligence approaches have been criticized as a "black box" ill-suited to aid human decision making. In this investigation, we assess a Human-Centered Artificial Intelligence (HCAI) approach to assessing personal qualities from text. Human raters coded 3,131 applicant essays describing out-of-school extracurricular and work experiences for seven different personal qualities. A pre-trained language model fine-tuned on this data successfully reproduced human codes and did so equally well across demographic subgroups. 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.

The importance of personal qualities for life outcomes is well established. Whether referred to as non-cognitive skills, social-emotional competencies, or character, personal qualities predict health, happiness, and success, even when controlling for cognitive ability and family socioeconomic status[1,2]. Likewise, personal qualities—usually measured using self-report questionnaires in low-stakes settings—predict both performance and persistence in college[3]. At present, college admissions consider personal qualities using a holistic review process, which relies on the judgments of human admissions officers[4,5]. Could artificial intelligence (AI) advance the goals of holistic review? AI has revolutionized the way information can be processed at scale, but there are growing concerns about whether and how AI should be used for consequential decision-making. Human-Centered Artificial Intelligence (HCAI) is an approach to AI that emphasizes human input and control in the design of intelligent systems and prioritizes interpretability, fairness, and transparency[6]. In this investigation, we evaluated an HCAI approach to assessing personal qualities in a large, national cohort of (N = 309,594) college applications.

Although fairness is an explicit objective of holistic review[7], history teaches us that considering personal qualities does not guarantee a more equitable college admissions process. According to Karabel[8], in the 1920s, Columbia University began requiring a personal essay among other elements attesting to an applicant's good "character." Previously, when admissions decisions had been based primarily on entrance exams and a growing proportion of Columbia's entering class was Jewish, its dean openly acknowledged the intention to return the campus to a state more welcoming for "students who come from homes of refinement" (p. 87). Other Ivy League universities soon followed suit. Using holistic review for exclusionary, rather than inclusionary, motives was possible because judgments of character were entirely in the eye of the beholder (i.e., the admissions officer).

Although the aim of holistic admissions may be nobler today than a century ago, the process by which it is carried out seems much the same. College admissions officers now say they most often look for evidence of personal qualities in the personal essay[9]. Logically, the words applicants use to describe themselves would be an appropriate place to hunt for evidence of personal qualities. Indeed thousands of studies have shown that language reveals individual differences in emotion, thought, and behavior[10].

What, specifically, do admissions officers do when reading what applicants have written about themselves? Best practices for the holistic review of college applications recommend using rubrics identifying a set of high-priority personal qualities that correlate with student success and/or are aligned with the school's mission and values[7,11]. It also is widely agreed that such rubrics should include clear definitions and examples, and be completed by not one but multiple admissions officers, each of whom independently review an applicant's materials.

For the vast majority of colleges, however, the status quo of holistic review likely departs from recommended best practices. Admissions office budgets are not unlimited, and even at selective, well-resourced universities, the sheer quantity of applications affords admissions officers minutes, not hours, to review each one[5,12]. It therefore seems unlikely that most colleges are able to assign multiple readers to independently and methodically rate each application on a variety of personal qualities. To the extent that each admissions officer instead relies on their own intuitive summary judgment of an applicant's overall character, the process is both opaque to applicants and practically impossible to audit. Moreover, even if the individual judgments of admissions officers are completely unbiased, there is the problem of noise: "Humans are unreliable decision makers," Kahneman and colleagues[13] observe. "Their judgments are strongly influenced by irrelevant factors, such as their current mood, the time since their last meal, and the weather." For instance, admissions officers weigh academic attributes more heavily when reading files on cloudier days and weigh non-academic attributes more heavily on sunnier days[14].

Can AI improve the assessment of personal qualities in college admissions? Efficiency is perhaps the most obvious advantage of AI. Once trained, computer algorithms generate scores

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

Table 1. Personal qualities, coding criteria, and example text

Personal quality: Criteria for coding by human raters	Fictionalized example essay with relevant phrases in italics
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.*

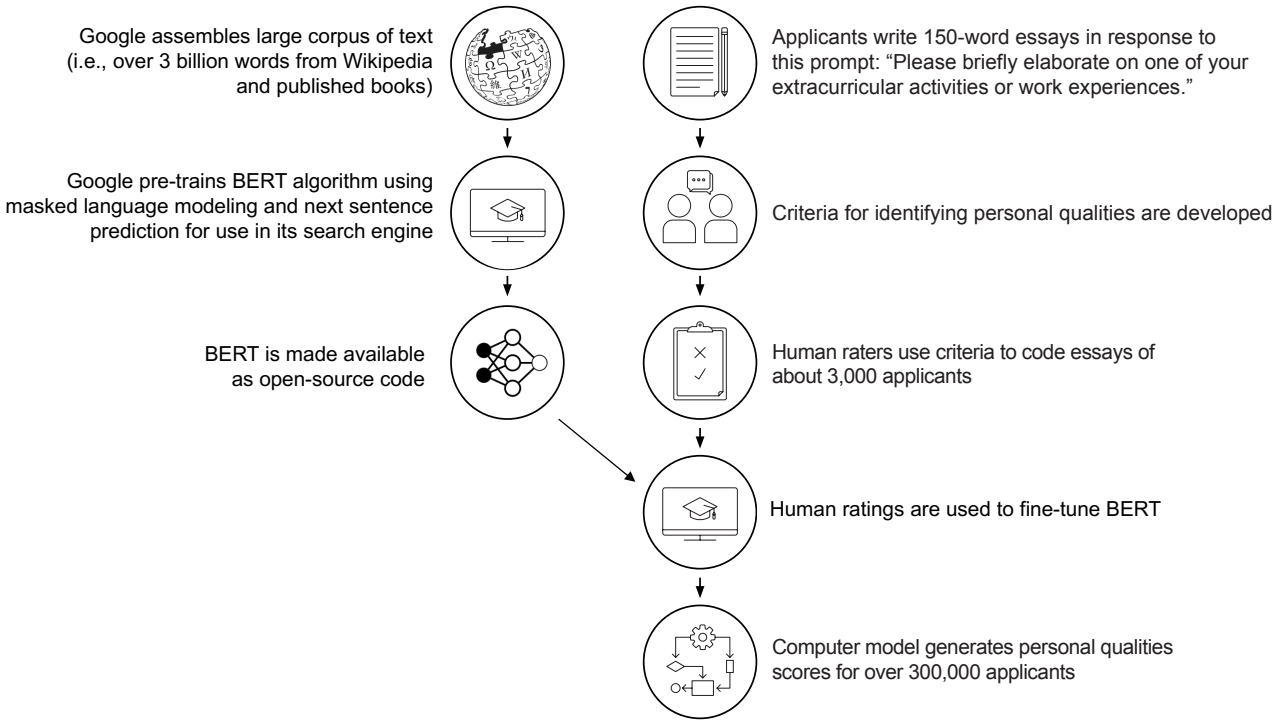


Fig. 1. A Human-Centered Artificial Intelligence (HCAI) approach to assessing personal qualities in college admissions

virtually instantaneously and at zero marginal cost for any number of applications. Algorithms also produce quantitative data which, unlike qualitative judgments, can be audited for evidence of bias, insofar as the data used for training aren't biased themselves and the algorithmic approach is interpretable. For instance, it is possible to test whether an algorithm is equally accurate across demographic subgroups, or whether the algorithm produces scores that differ systematically across subgroups in unexpected ways[15]. A final advantage is that algorithms are reliable. Unlike human beings, algorithms are not influenced by extraneous factors like hunger, fatigue, weather, or time of day.

Despite these advantages, AI has been criticized as a "black box" that can inadvertently introduce bias and harm underprivileged subgroups. For instance, AI has been shown to perpetuate, or even exacerbate, existing bias, in the domains of predictive policing, job hiring, and medical diagnosis[16–18]. When applied to college essays, structural topic modeling and dictionary algorithms, which operate in an unsupervised fashion (i.e., with little-to-no human input), have been shown to predict applicants' household income more than did SAT scores[19]. Even if an algorithm does not take as input protected personal information (e.g., gender or household income), it can indirectly use this information through its correlations with features it does observe (e.g., zip code as a proxy for socioeconomic status).

In recent years, there has been increasing interest in remedying unfairness and discrimination in algorithms[20,21]. Whereas several of the approaches focus at the algorithmic level (i.e., algorithmic fairness), this alone is insufficient when it comes to predicting personal qualities. In particular, because these are psychological constructs (e.g., persistence)—which can only be inferred rather than directly measured the way, for example, recidivism can—they require human judgments to ensure that the resultant measures are both accurate as well as unbiased[22]. Accordingly, we suggest that Human-Centered Artificial Intelligence (HCAI) can provide a promising approach to developing assessments that go beyond predictive accuracy to include fairness, transparency, and interpretability[6]. Although HCAI is a broad and emerging research area, a key tenet is that humans (not AI) are placed in the center of how intelligent systems are designed and used—akin to placing the sun instead of earth in the center of the solar system[23]. In the present case, this approach entails leveraging human judgments in the design and use of AI-based assessments of personal qualities.

In this investigation, we capitalized on a national, longitudinal sample of 309,594 college applications to evaluate an HCAI approach to assessing personal qualities (see **Figure 1**). Application data were matched with six-year college graduation data and de-identified prior to our analysis (See **Supplementary Materials** for details). Each application included a 150-word essay in which students elaborated on extracurricular and work experiences outside of school. Research assistants, who were blind to all other aspects of the application, used the criteria in **Table 1** to identify seven personal qualities in a *Development Sample* of 3,131 essays. We then used this labeled dataset to fine-tune an open-source language representation model called BERT[24]. We evaluated evidence of convergent and discriminant validity of this model, as well as how its predictions were related to demographic characteristics (i.e., bias). We then used the model to produce computer-generated likelihoods of personal qualities in the *Holdout Sample* of 306,463 essays. Finally, we examined the patterns of association between computer-generated

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)	**.83*****	−.01	−.04*	−.09***	−.11***	−.05**	.05**
2. Leadership (LD)	.00	**.79*****	.14***	−.02	.00	−.09***	.05*
3. Teamwork (TW)	−.05**	.19***	**.54*****	.07***	.08***	−.01	.06**
4. Learning (LR)	−.09***	−.04*	.03	**.75*****	.09***	−.02	−.02
5. Perseverance (PS)	−.15***	−.03	.05*	.09***	**.63*****	.06**	.04*
6. Intrinsic Motivation (IM)	−.05**	−.10***	.00	−.01	.05**	**.70*****	.02
7. Goal Pursuit (GP)	.08***	.08***	.07***	−.02	.04*	.03	**.55*****
Descriptive Statistics							
Human Interrater 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	.27	.45	.20	.47	.34

likelihoods and demographic characteristics and, as evidence of criterion validity, whether they predicted six-year college graduation beyond a rich set of covariates.

Results

Human raters identified an average of two personal qualities in each essay in the Development Sample, but some personal qualities were more commonly observed than others. For instance, the personal qualities of learning and intrinsic motivation were identified in 42% of essays, whereas the personal qualities of leadership and perseverance were coded for only 18% and 19% of essays, respectively. Using these human ratings, we fine-tuned a Bidirectional Encoder Representations from Transformers (BERT) model. This resulted in seven computer-generated, continuous (0 to 1) personality quality scores for each essay. See **Supplementary Materials** for details.

Convergent and discriminant validity of computer-generated likelihoods in the Development Sample. As evidence of convergent validity, computer-generated likelihoods for each personal quality converged with human ratings of the same personal quality (rs ranged from .54 to .83, average $r = .70$). As evidence of discriminant validity, computer-generated likelihoods for a particular personal quality did not correlate with human ratings of other personal qualities, and vice versa (rs from -.15 to .19, 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 ($r = .93$, $p < .001$).

Convergent validity does not vary by demographic subgroup in the Development Sample. As shown in **Table 3,** correlations 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 .77 for female applicants and .78 for male applicants. As shown in **Table S6,** 9% of the correlations differed by subgroup. In half of these comparisons, the model was more accurate for the marginalized group, while in the other half, the majority subgroup was favored.

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[25]. See **Table S3** in **Supplementary Materials** for details.

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 algorithm to the *Holdout Sample* of 306,463 essays. Again, 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 $|\mathrm{d}| = 0.17$). 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 modestly predictive of college graduation (ORs from 1.025 to 1.135, ps $< .001$; $AUC = .562$). As shown in Model 2 in **Table 4,** five of seven personal qualities remained predictive of college graduation when controlling for demographics, standardized test scores, and out-of-school activities (ORs from 1.015 to 1.075, ps $< .002$). See **Supplementary Materials** for details on imputation and robustness checks.

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 a large language model using human ratings of prosocial purpose, leadership, teamwork, learning, perseverance, intrinsic motivation, and goal pursuit in applicants' essays about their out-of-school activities. This algorithm demonstrated convergent and discriminant validity with human ratings of same and different personal qualities, respectively, and these patterns were consistent across demographic

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	0.84	0.78	0.49	0.79	0.61	0.70	0.57	0.78	−0.01	−0.16	0.14
Black	487	0.80	0.74	0.62	0.76	0.70	0.71	0.53	0.77	0.00	−0.26	0.21
Latino	501	0.82	0.83	0.55	0.70	0.60	0.69	0.60	0.80	0.03	−0.12	0.27
Asian	590	0.82	0.77	0.52	0.72	0.64	0.66	0.51	0.75	0.01	−0.13	0.16
Other	290	0.83	0.80	0.56	0.72	0.60	0.75	0.53	0.77	0.01	−0.18	0.24
No race reported	369	0.88	0.81	0.59	0.75	0.65	0.70	0.58	0.79	0.02	−0.15	0.23
Number of parents with college degrees												
None	1608	0.81	0.78	0.54	0.75	0.63	0.73	0.57	0.79	0.01	−0.17	0.23
One	563	0.83	0.80	0.55	0.77	0.65	0.64	0.51	0.76	0.01	−0.15	0.13
Two	853	0.86	0.78	0.54	0.72	0.62	0.67	0.54	0.77	0.00	−0.12	0.17
Gender												
Female	1702	0.83	0.78	0.55	0.74	0.62	0.69	0.55	0.77	0.01	−0.18	0.21
Male	1413	0.82	0.79	0.53	0.75	0.64	0.71	0.56	0.78	0.00	−0.13	0.17
Married parents												
Parents married	2055	0.83	0.79	0.53	0.75	0.62	0.70	0.55	0.78	0.01	−0.14	0.18
Parents not married	1061	0.84	0.77	0.58	0.75	0.64	0.69	0.55	0.77	0.01	−0.16	0.21
English language learner status												
English language learner	808	0.85	0.74	0.48	0.72	0.62	0.71	0.54	0.77	0.01	−0.13	0.20
Native speaker	2308	0.82	0.80	0.57	0.75	0.63	0.69	0.56	0.78	0.01	−0.16	0.19
Title 1 status of high school												
Title 1 public school	1127	0.81	0.81	0.56	0.74	0.65	0.72	0.56	0.78	0.01	−0.20	0.25
Non-Title 1 school	1552	0.85	0.78	0.55	0.76	0.61	0.68	0.56	0.78	0.01	−0.14	0.17

subgroups. As further evidence of validity, computer-generated personal quality scores prospectively predicted college graduation six years later.

We found computer-generated scores to be fairly independent of demographics and less correlated with demographics than either standardized test scores or out-of-school activities. In contrast, Alvero and colleagues[19] found that certain features extracted from college application essays were as related to household income as SAT scores. For instance, the topic "seeking answers," indicated by words like "question" and "book," correlated with household income at $r = .28$. Likewise, Pennebaker and colleagues[26] found that categorical versus dynamic words in college essays correlated with parental education at $r = .22$. Why do our results differ? We cannot know for certain. Alvero et al.[19] used essays from the University of California system, and Pennebaker et al.[26] 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. But more critically, rather than using unsupervised topic modeling or dictionary approaches, we used a supervised machine learning approach in which we fine-tuned a language representation model to match human ratings of personal qualities. Finally, it is possible 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.

It is worth noting that 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. As context, a growing literature suggests that life outcomes are generally extremely difficult to predict with precision[27]. One reason is that the more numerous the influences on behavior, the smaller the influence of any single factor[28]. Graduating from college more than six years after students turn in their applications depends on myriad factors other than the personal qualities we identified in this study. For instance, students may drop out because they cannot afford tuition payments[29], because they are not prepared academically[30,31], or because they feel like they do not belong[32]. Regardless, it is almost certain that we would have obtained a stronger signal of personal qualities had we had access to additional sources of information (e.g., the text of letters of recommendation from teachers and guidance counselors). Indeed, the underlying "more is more" principle of holistic review (i.e., having multiple raters read each application) is consistent with the psychometric principle of aggregation[33] and the demonstrated superior predictive validity of composite over component measures[1,34,35].

Several limitations of our investigation are worth highlighting.

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 4. 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	1.135***	1.076***
	(0.005)	(0.005)
Leadership	1.135***	1.066***
	(0.005)	(0.005)
Teamwork	1.094***	1.044***
	(0.005)	(0.005)
Learning	1.065***	1.044***
	(0.004)	(0.005)
Perseverance	1.079***	1.015**
	(0.005)	(0.005)
Intrinsic Motivation	1.061***	1.005
	(0.004)	(0.005)
Goal Pursuit	1.025***	1.005
	(0.004)	(0.005)
Race/ethnicity (vs. White)		
Black		0.775***
		(0.019)
Latino		0.870***
		(0.019)
Asian		0.735***
		(0.017)
Other		0.750***
		(0.017)
No race reported		0.850***
		(0.013)
Parental education (vs. No parent w/ college degree)		
One parent w/ college degree		1.198***
		(0.012)
Two parents w/ college degree		1.334***
		(0.012)
Female		1.434***
		(0.010)
Married parents		1.310***
		(0.011)
English language learner		0.768***
		(0.015)
Title 1 high school		0.950***
		(0.013)
Out-of-school activities (OSA)		
Number of OSA		1.247***
		(0.005)
Time per OSA		1.088***
		(0.004)
Proportion sports		1.039***
		(0.005)
Standardized test scores		1.488***
		(0.006)
Constant	3.558***	2.536***
	(0.004)	(0.014)
AUC	.562	.690

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

Table S25) affirm the conclusions of our main analyses, future research should not face this limitation.

Third, the only outcome for which we could evaluate the predictive validity of our algorithm was college graduation. Future research should include other important outcomes, such as academic performance, extracurricular involvement, mental health, and contributions to the campus community[36].

Finally, the interrater 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 motivation) 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 algorithm depends on the quality of the training data[37]. We speculate that as subject matter experts, college admissions officers would in future research, and in practice, produce more reliable ratings than the trained research assistants in our investigation.

In our view, a Human-Centered Artificial Intelligence approach to measuring personal qualities warrants both optimism and caution. This investigation demonstrates that algorithms trained on human ratings can be efficient (yielding millions of personal quality scores in a matter of minutes) and fair to demographic subgroups. Moreover, as suggested by the criteria in **Table 1,** this approach produces scores that directly index personal qualities. When put into practice, however, there is the possibility of algorithm aversion—the tendency to trust human decision makers over algorithms even in the face of contradictory evidence[38]. Even more concerning is Campbell's Law[39], which states that the more a given measure matters for high-stakes decisions (as opposed to research), the greater the incentive to distort it. It is not hard to imagine how applicants might try to game the system, shaping their essays to match what admissions officers, and the algorithms they train, are looking for. Moreover, applicants from more advantaged backgrounds may be better positioned to do so. Nevertheless, progress depends on dissatisfaction with the status quo, and when it comes to 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 [40]. 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.

BERT fine tuning procedure. Bidirectional Encoder Representations from Transformers (BERT) [24] 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). This pre-training allows BERT to obtain better accuracy than other models trained from scratch. 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). See **Supplementary Materials** for technical details on our fine-tuning process.

To begin, 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 these manually annotated datasets to fine-tune a BERT model 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 holdout sample of 306,463 essays, yielding more than two million continuous codes.

1. TE Moffitt, et al., A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* **108**, 2693–2698 (2011).
2. M Almlund, AL Duckworth, J Heckman, T Kautz, Personality Psychology and Economics in *Handbook of the Economics of Education.* (Elsevier) Vol. 4, pp. 1–181 (2011).
3. SB Robbins, et al., Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychol. Bull.* **130**, 261–288 (2004).
4. A Alvero, et al., AI and Holistic Review: Informing Human Reading in College Admissions in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.* (ACM, New York NY USA), pp. 200–206 (2020).
5. E Hoover, Working Smarter, Not Harder, in Admissions (2017) Section: News.
6. MO Riedl, Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **1**, 33–36 (2019).
7. AL Coleman, JL Keith, Understanding Holistic Review in Higher Education Admissions, (College Board, New York), Technical report (2018).
8. J Karabel, *The chosen: The hidden history of admission and exclusion at Harvard, Yale, and Princeton.* (Houghton Mifflin Harcourt), (2005).
9. National Association for College Admission Counseling, Character and the college admission process, Technical report (2020).
10. RL Boyd, JW Pennebaker, Language-based personality: a new approach to personality in a digital world. *Curr. Opin. Behav. Sci.* **18**, 63–68 (2017).
11. TR Anderson, R Weissbourd, Character Assessment in College Admission, (Making Caring Common Project, Boston), Technical report (2020).
12. M Korn, Some Elite Colleges Review an Application in 8 Minutes (or Less). *Wall Str. J.* (2018).
13. D Kahneman, AM Rosenfield, L Gandhi, T Blaser, Noise. How to Overcome the High, Hidden Cost of Inconsistent Decision Making. *Harv. Bus. Rev.* pp. 38–46 (2016).
14. U Simonsohn, Clouds make nerds look good: field evidence of the impact of incidental factors on decision making. *J. Behav. Decis. Mak.* p. 10 (2006).
15. L Tay, SE Woo, L Hickman, BM Booth, S D'Mello, A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Adv. Methods Pract. Psychol. Sci.* **5**, 1–30 (2022).
16. J Manyika, J Silberg, B Presten, What Do We Do About the Biases in AI? p. 5 (2019).
17. Z Obermeyer, B Powers, C Vogeli, S Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. p. 8 (2019).
18. D Ensign, SA Friedler, S Neville, C Scheidegger, S Venkatasubramanian, Runaway Feedback Loops in Predictive Policing. *Proc. Mach. Learn. Res.* **81**, 12 (2018).
19. A Alvero, et al., Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Sci. Adv.* **7**, eabi9031 (2021).
20. S Verma, J Rubin, Fairness definitions explained in *Proceedings of the International Workshop on Software Fairness.* (ACM, Gothenburg Sweden), pp. 1–7 (2018).
21. K Holstein, J Wortman Vaughan, H Daumé, M Dudik, H Wallach, Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* (ACM, Glasgow Scotland Uk), pp. 1–16 (2019).
22. SK D'Mello, L Tay, R Southwell, Psychological Measurement in the Information Age: Machine-Learned Computational Models. *Curr. Dir. Psychol. Sci.* **31**, 76–87 (2022).
23. B Shneiderman, Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interact.* pp. 109–124 (2020).

24. J Devlin, MW Chang, K Lee, K Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Arxiv Prepr.* **1810.04805v2** (2019).
25. L Kamas, A Preston, Empathy, gender, and prosocial behavior. *J. Behav. Exp. Econ.* **92**, 101654 (2021).
26. JW Pennebaker, CK Chung, J Frazee, GM Lavergne, DI Beaver, When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLoS ONE* **9**, e115844 (2014).
27. MJ Salganik, et al., Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci.* **117**, 8398–8403 (2020).
28. S Ahadi, E Diener, Multiple determinants and effect size. *J. Pers. Soc. Psychol.* **56**, 398–406 (1989).
29. S Goldrick-Rab, Following Their Every Move: An Investigation of Social-Class Differences in College Pathways. *Sociol. Educ.* **79**, 67–79 (2006).
30. D Hepworth, B Littlepage, K Hancock, Factors influencing university student academic success. *Educ. Res. Q.* **42**, 45–61 (2018).
31. SF Porchea, J Allen, S Robbins, RP Phelps, Predictors of Long-Term Enrollment and Degree Outcomes for Community College Students: Integrating Academic, Psychosocial, Sociodemographic, and Situational Factors. *The J. High. Educ.* **81**, 680–708 (2010).
32. MC Murphy, et al., A customized belonging intervention improves retention of socially disadvantaged students at a broad-access university. *Sci. Adv.* **6**, eaba4677 (2020).
33. JP Rushton, CJ Brainerd, M Pressley, Behavioral development and construct validity: The principle of aggregation. *Psychol. bulletin* **94**, 18 (1983).
34. DJ Benjamin, et al., 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).
35. AL Duckworth, ME Seligman, Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychol. Sci.* **16**, 939–944 (2005).
36. WW Willingham, *Success in college: The role of personal qualities and academic ability*. (College Board Publications), (1985).
37. RS Geiger, et al., "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.
38. BJ Dietvorst, JP Simmons, C Massey, Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *J. Exp. Psychol. Gen.* **144**, 114–126 (2015).
39. DT Campbell, Assessing the impact of planned social change. *Eval. Program Plan.* **2**, 67–90 (1979).
40. 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*], Angela L. Duckworth[1], Margo Gardner[2], Abigail Quirk[1], Cathlyn Stone[2], Arjun Rao[2], Stephen Hutt[1], and Sidney K. D'Mello[2]

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*).

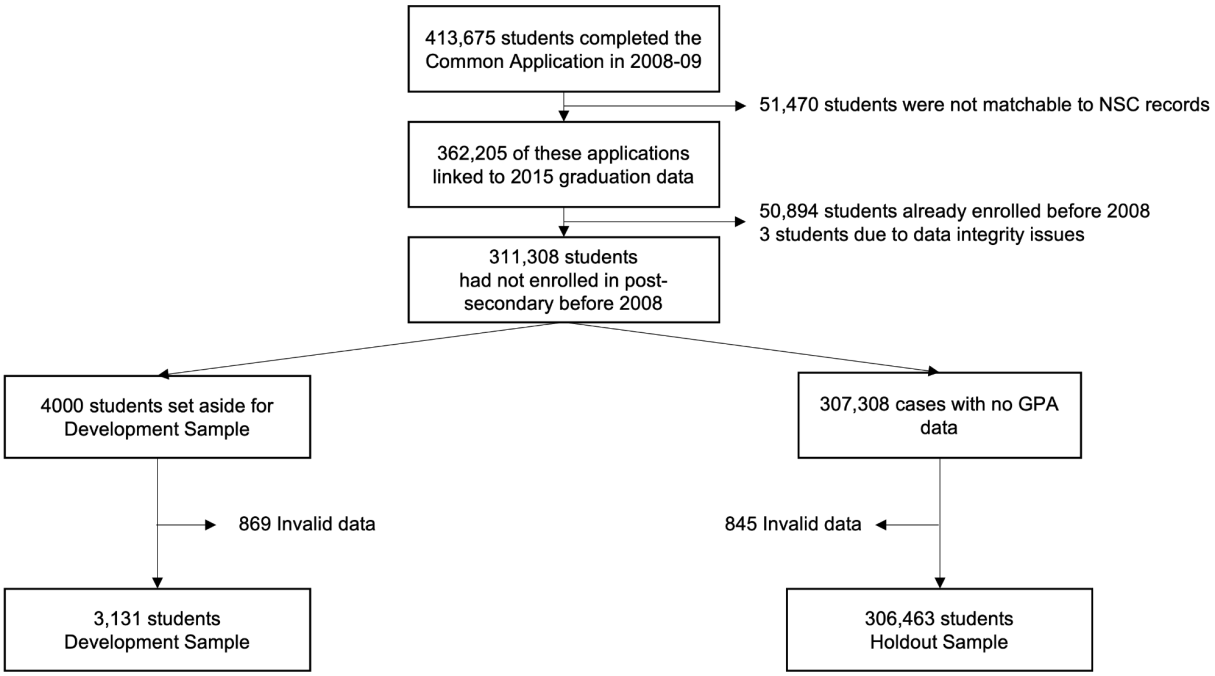
Specifically, our sample was drawn from the population of 413,675 students who completed the Common App during the 2008-09 academic year for college admission during the 2009-10 academic year. From this population, we selected the 311,308 students who had not enrolled in a postsecondary institution prior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

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



BERT algorithm fine-tuning procedure

We used the BERT-base-uncased model, which we obtained from huggingface's "transformers" Python library. See this link to the model hosted on the huggingface website. We used 4 training epochs, with 32 examples used to predict on before updating the weights in each iteration. 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 BERT 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 BERT procedure entails applying these 70 BERT 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	.00											
3. Teamwork	-.06**	.20***										
4. Learning	-.09***	-.03	.05**									
5. Perseverance	-.15***	-.03	.04*	.10***								
6. Intrinsic motivation	-.06***	-.13***	.00	-.02	.07***							
7. Goal pursuit	.09***	.08***	.07***	-.03†	.04*	.04*						
8. Standardized test scores	-.01	.08***	.07***	.04†	.10***	.02	.02					
9. Number of activities	.08***	.14***	.12***	.08***	.06***	-.01	.09***	.36***				
10. Time per activity	-.01	.13***	.06***	.04*	.05**	.03†	.06**	.26***	.40***			
11. Proportion sports	-.13***	.02	.05**	.02	.05**	.06***	.02	-.11***	-.06**	.28***		
12. College graduation	.03	.07***	.04*	.06***	.03	.03	.04*	.30***	.22***	.17***	.00	
M	0.36	0.19	0.27	0.45	0.2	0.47	0.34	1,693	3.46	2.11	23.34%	66.3%
SD	0.46	0.37	0.34	0.45	0.34	0.45	0.37	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. *** $p < .001$. ** $p < .01$. * $p < .05$.

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	-.12***	.26***										
4. Learning	-.13***	-.02***	.05***									
5. Perseverance	-.20***	-.04***	.06***	.11***								
6. Intrinsic motivation	-.09***	-.14***	-.04***	-.03***	.08***							
7. Goal pursuit	.11***	.07***	.00†	-.08***	.01***	.01***						
8. Standardized test scores	-.01***	.07***	.06***	.00	.08***	.01***	.02***					
9. Number of activities	.10***	.09***	.08***	.04***	.05***	.02***	.08***	.34***				
10. Time per activity	-.03***	.07***	.02***	.00†	.02***	.06***	-.01***	.14***	.03***			
11. Proportion sports	-.10***	-.03***	.03***	.01***	.08***	.04***	-.03***	-.21***	-.27***	.09***		
12. College graduation	.04***	.05***	.05***	.02***	.03***	.01***	.02***	.22***	.18***	.09***	-.05***	
M	0.34	0.21	0.33	0.49	0.24	0.56	0.42	1,826	5.16	2.53	0.26	0.78
SD	0.44	0.34	0.3	0.43	0.32	0.42	0.33	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. Values are Matthews 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	−0.09	0.09	0.02	−0.01	0.06	0.11	0.01
Black	0.00	−0.05	−0.13	−0.08	−0.17	−0.09	0.06
Latino	0.10	−0.03	0.03	−0.05	0.03	0.00	−0.05
Asian	0.11	−0.05	0.08	0.11	0.06	−0.13	0.01
Other	−0.02	0.09	−0.07	−0.11	−0.12	0.00	0.08
Missing	−0.11	−0.08	0.02	0.11	0.07	0.10	−0.11
Number of parents with college degrees							
None	0.08	−0.03	−0.11	−0.09	−0.08	0.00	−0.03
One	−0.05	0.01	0.05	−0.02	−0.06	0.00	−0.03
Two	−0.06	0.03	0.10	0.14	0.14	0.01	0.06
Female	0.26	0.09	0.04	0.09	0.07	0.11	0.08
Married parents	−0.02	0.04	0.08	0.04	0.07	0.00	0.06
English language learner	0.11	−0.07	0.04	0.01	0.02	−0.07	0.04
Title 1 High School	0.00	0.02	0.01	−0.06	−0.05	0.04	−0.01

Note. 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	0.06	0.02	0.00	0.07	−0.02
Black	0.01	−0.03	−0.12	−0.11	−0.16	−0.14	−0.04
Latino	0.08	−0.04	−0.08	−0.01	−0.07	−0.06	−0.01
Asian	0.07	0.00	−0.01	0.08	0.09	−0.11	0.07
Other	0.00	−0.01	−0.05	−0.03	−0.03	0.00	0.02
Missing	−0.03	−0.03	0.01	−0.02	0.04	0.03	0.00
Number of parents with college degrees							
None	0.00	−0.07	−0.09	−0.05	−0.11	−0.08	−0.06
One	−0.01	−0.01	−0.01	0.00	−0.03	0.01	0.00
Two	0.00	0.06	0.08	0.04	0.10	0.06	0.05
Female	0.22	0.06	0.01	0.04	0.00	0.12	0.04
Married parents	0.05	0.06	0.06	0.04	0.05	0.02	0.03
English language learner	0.07	−0.06	−0.06	0.04	0.04	−0.10	0.07
Title 1 high school	−0.02	0.01	−0.01	0.00	−0.04	−0.06	−0.02

Note. 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	.80	.84	-.04	.74	.78	-.04	.62	.49	.13**	.76	.79	-.03	.70	.61	.09	.71	.70	.01	.53	.57	-.03
Latino	.82	.84	-.02	.83	.78	.05	.55	.49	.06	.70	.79	-.09**	.60	.61	-.01	.69	.70	-.01	.60	.57	.03
Asian	.82	.84	-.03	.77	.78	-.01	.52	.49	.03	.72	.79	-.06	.64	.61	.02	.66	.70	-.04	.51	.57	-.06
Other	.83	.84	-.01	.80	.78	.02	.56	.49	.07	.72	.79	-.07	.60	.61	-.01	.75	.70	.05	.53	.57	-.04
No race reported	.88	.84	.04	.81	.78	.03	.59	.49	.10	.75	.79	-.03	.65	.61	.04	.70	.70	-.01	.58	.57	.01
Number of parents with college degrees (vs. None)																					
One	.83	.81	.02	.80	.78	.02	.55	.54	.01	.77	.75	.02	.65	.63	.02	.64	.73	-.09**	.51	.57	-.06
Two	.86	.81	.05**	.78	.78	-.00	.54	.54	.00	.72	.75	-.02	.62	.63	-.01	.67	.73	-.06*	.54	.57	-.03
Other demographics																					
Male	.82	.83	-.01	.79	.78	.00	.53	.55	-.02	.75	.74	.01	.64	.62	.02	.71	.69	.02	.56	.55	.01
Parents not married	.84	.83	.01	.77	.79	-.02	.58	.53	.05	.75	.75	-.00	.64	.62	.02	.69	.70	-.01	.55	.55	-.01
English language learner	.85	.82	.02	.74	.80	-.06**	.48	.57	-.09*	.72	.75	-.03	.62	.63	-.01	.71	.69	.01	.54	.56	-.01
Title 1 public school	.81	.85	-.03*	.81	.78	.02	.56	.55	.01	.74	.76	-.02	.65	.61	.04	.72	.68	.04	.56	.56	-.00

Note. Values are 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	.84***	-.03	-.09**	-.09*	-.11**	-.06	.09*
2. Leadership	-.02	.78***	.07	-.03	-.05	-.13***	.01
3. Teamwork	-.05	.14***	.49***	.08*	.07*	.02	.02
4. Learning	-.11***	-.05	.04	.79***	.11**	-.05	-.06
5. Perseverance	-.15***	-.09**	.02	.08*	.61***	.04	.04
6. Intrinsic motivation	-.02	-.16***	.04	-.02	.07*	.70***	.04
7. Goal pursuit	.10**	-.01	.00	-.03	.03	.03	.57***
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	0.45	0.21	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	.80***	.00	.01	-.05	-.20***	-.06	.01
2. Leadership	.04	.74***	.18***	-.07	-.10*	-.06	.05
3. Teamwork	.03	.21***	.62***	.06	.03	.00	.03
4. Learning	-.06	-.05	.03	.76***	.07	-.02	-.03
5. Perseverance	-.26***	-.07	.05	.05	.70***	.05	-.04
6. Intrinsic motivation	-.10*	-.05	.00	-.02	.01	.71***	.11*
7. Goal pursuit	.00	.04	.12**	.00	.04	.04	.53***
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	0.36

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	.82***	.00	.00	-.10*	-.12**	-.04	.07
2. Leadership	-.01	.83***	.19***	-.03	.11*	.00	.09*
3. Teamwork	-.11*	.27***	.55***	.08	.12**	.01	.11*
4. Learning	-.11*	-.03	.09	.70***	.10*	-.02	-.02
5. Perseverance	-.12**	.07	.10*	.06	.60***	.05	.04
6. Intrinsic motivation	-.03	-.05	-.09*	.03	.05	.69***	-.05
7. Goal pursuit	.06	.15***	.10*	-.05	.13**	.04	.60***
Frequency of human rating	0.37	0.18	0.27	0.41	0.22	0.41	0.28
Mean of computer-generated likelihood	0.40	0.18	0.28	0.43	0.21	0.47	0.33

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	.82***	.01	-.12**	-.12**	-.07	-.03	.03
2. Leadership	.00	.77***	.16***	.00	.04	-.13**	.03
3. Teamwork	-.11**	.14***	.52***	.08	.07	-.03	.04
4. Learning	-.07	-.08	-.02	.72***	.01	-.02	.00
5. Perseverance	-.10*	.00	.04	.06	.64***	.12**	.10*
6. Intrinsic motivation	-.06	-.08*	.01	-.01	.15***	.66***	-.06
7. Goal pursuit	.07	.16***	.14***	.00	.07	.01	.51***
Frequency of human rating	0.40	0.16	0.30	0.47	0.22	0.38	0.32
Mean of computer-generated likelihood	0.40	0.18	0.29	0.49	0.21	0.42	0.34

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	.83***	.01	-.03	-.17**	-.18**	.00	-.03
2. Leadership	.01	.80***	.16**	.06	.01	-.18**	.12*
3. Teamwork	-.07	.24***	.56***	.03	.10	-.07	.11
4. Learning	-.09	.00	.03	.72***	.13*	.06	.13*
5. Perseverance	-.14*	-.09	.07	.11	.60***	-.05	.02
6. Intrinsic motivation	.08	-.18**	-.08	.10	-.03	.75***	.07
7. Goal pursuit	.09	.10	.05	.11	-.06	.03	.53***
Frequency of human rating	0.36	0.20	0.20	0.35	0.14	0.43	0.34
Mean of computer-generated likelihood	0.35	0.22	0.25	0.41	0.16	0.47	0.37

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	.88***	-.03	.03	-.04	-.07	-.11*	.09
2. Leadership	.03	.81***	.17***	.05	-.01	-.05	.00
3. Teamwork	.04	.23***	.59***	.06	.09	-.01	.09
4. Learning	-.07	.05	.02	.75***	.16**	-.01	-.06
5. Perseverance	-.15**	.02	-.02	.16**	.65***	.07	.04
6. Intrinsic motivation	-.12*	-.01	.04	-.06	-.02	.70***	.05
7. Goal pursuit	.14**	.10*	.07	-.05	.02	.03	.58***
Frequency of human rating	0.30	0.15	0.26	0.43	0.20	0.43	0.26
Mean of computer-generated likelihood	0.32	0.17	0.28	0.50	0.22	0.51	0.31

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	.81***	-.03	-.03	-.07**	-.12***	-.03	.05*
2. Leadership	-.03	.78***	.16***	.02	.01	-.08**	.06*
3. Teamwork	-.07**	.23***	.54***	.10***	.05*	-.01	.04
4. Learning	-.08**	-.03	.07**	.75***	.06*	-.04	-.01
5. Perseverance	-.17***	.01	.04	.07**	.63***	.03	.01
6. Intrinsic motivation	-.02	-.07**	-.02	-.01	.04	.73***	.02
7. Goal pursuit	.06*	.11***	.07**	-.01	.06*	.06*	.57***
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	0.25	0.43	0.19	0.47	0.34

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	.83***	.05	-.06	-.09*	-.08*	-.04	.04
2. Leadership	.05	.80***	.05	-.05	-.04	-.12**	.04
3. Teamwork	-.03	.13***	.55***	.05	.13**	.04	.09*
4. Learning	-.07	-.03	.05	.77***	.13**	.03	-.03
5. Perseverance	-.13***	-.11**	.06	.13***	.65***	.01	.02
6. Intrinsic motivation	-.04	-.15***	.08*	-.02	.07	.64***	.03
7. Goal pursuit	.10**	.03	.04	.00	-.03	-.03	.51***
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	0.47	0.33

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	.86***	-.01	-.06	-.12***	-.12***	-.10**	.06
2. Leadership	.01	.78***	.17***	-.06	.00	-.11**	.02
3. Teamwork	-.02	.17***	.54***	.02	.10**	-.05	.06
4. Learning	-.12***	-.07	-.05	.72***	.12***	-.03	-.04
5. Perseverance	-.12***	-.06	.04	.08*	.62***	.13***	.09**
6. Intrinsic motivation	-.11**	-.11***	-.03	.00	.07	.67***	.01
7. Goal pursuit	.08*	.04	.11**	-.05	.07*	.01	.54***
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	0.29	0.50	0.23	0.47	0.36

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	.83***	-.01	-.06*	-.10***	-.15***	-.02	.06*
2. Leadership	.00	.78***	.13***	-.01	.02	-.08***	.03
3. Teamwork	-.07**	.21***	.55***	.06**	.10***	-.02	.04
4. Learning	-.09***	-.03	.06*	.74***	.09***	-.03	-.01
5. Perseverance	-.18***	-.01	.05	.09***	.62***	.06**	.03
6. Intrinsic motivation	-.03	-.11***	-.01	-.02	.04	.69***	.01
7. Goal pursuit	.08***	.07**	.09***	-.01	.04	.03	.55***
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	0.21	0.28	0.47	0.21	0.49	0.35

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	.82***	-.02	-.04	-.09***	-.08**	-.11***	.04
2. Leadership	-.02	.79***	.14***	-.03	-.04	-.12***	.07**
3. Teamwork	-.04	.17***	.53***	.08**	.06*	.01	.07**
4. Learning	-.11***	-.05	.00	.75***	.09***	-.01	-.03
5. Perseverance	-.13***	-.06*	.04	.07**	.64***	.04	.05
6. Intrinsic motivation	-.10***	-.08**	.00	.00	.07*	.71***	.04
7. Goal pursuit	.06*	.08**	.04	-.02	.04	.03	.56***
Frequency of human rating	0.27	0.17	0.24	0.39	0.17	0.39	0.32
Mean of computer-generated likelihood	0.30	0.17	0.26	0.43	0.18	0.44	0.33

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	.83***	-.02	-.05*	-.11***	-.12***	-.05*	.05*
2. Leadership	.00	.79***	.14***	-.01	.01	-.09***	.04
3. Teamwork	-.06**	.18***	.53***	.07**	.09***	.01	.06**
4. Learning	-.10***	-.05*	.02	.75***	.10***	-.03	-.03
5. Perseverance	-.14***	-.03	.04	.08***	.62***	.05*	.05*
6. Intrinsic motivation	-.05*	-.11***	.01	-.01	.04	.70***	.01
7. Goal pursuit	.07**	.09***	.08***	-.03	.04*	.01	.55***
Frequency of human rating	0.34	0.18	0.26	0.42	0.19	0.42	0.32
Mean of computer-generated likelihood	0.36	0.20	0.28	0.46	0.21	0.47	0.35

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	.84***	.01	-.03	-.05	-.11***	-.07*	.05
2. Leadership	.00	.77***	.14***	-.02	-.02	-.11***	.05
3. Teamwork	-.04	.21***	.58***	.09**	.07*	-.03	.04
4. Learning	-.07*	-.01	.07*	.75***	.07*	.00	.01
5. Perseverance	-.16***	-.04	.05	.09**	.64***	.07*	.01
6. Intrinsic motivation	-.04	-.07*	-.03	.00	.09**	.69***	.04
7. Goal pursuit	.09**	.06	.06*	.01	.04	.06*	.55***
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	0.44	0.18	0.47	0.33

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	.85***	.03	-.05	-.13***	-.10**	-.01	.01
2. Leadership	.04	.74***	.12***	-.01	.03	-.07*	.04
3. Teamwork	-.07*	.20***	.48***	.07*	.06	-.02	.07*
4. Learning	-.13***	-.06	.04	.72***	.05	-.01	-.01
5. Perseverance	-.13***	-.02	.04	.06	.62***	.05	.08*
6. Intrinsic motivation	-.02	-.09*	.01	.02	.05	.71***	-.04
7. Goal pursuit	.07*	.14***	.09**	.01	.06	.04	.54***
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	0.28	0.46	0.20	0.45	0.35

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	.82***	-.02	-.05*	-.08***	-.12***	-.07**	.06**
2. Leadership	-.01	.80***	.15***	-.02	-.01	-.10***	.05*
3. Teamwork	-.05*	.19***	.57***	.07***	.09***	.00	.05*
4. Learning	-.08***	-.03	.03	.75***	.11***	-.02	-.02
5. Perseverance	-.16***	-.03	.05*	.09***	.63***	.06**	.02
6. Intrinsic motivation	-.06**	-.10***	-.01	-.02	.06**	.69***	.04*
7. Goal pursuit	.08***	.06**	.06**	-.02	.04	.03	.56***
Frequency of human rating	0.33	0.18	0.25	0.41	0.18	0.43	0.31
Mean of computer-generated likelihood	0.35	0.20	0.26	0.45	0.20	0.48	0.34

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	.81***	-.02	-.05	-.06*	-.17***	-.05	.05
2. Leadership	.01	.81***	.18***	-.03	.04	-.06*	.07*
3. Teamwork	-.05	.25***	.56***	.06*	.13***	-.02	.06
4. Learning	-.04	-.05	.03	.74***	.10***	-.03	-.05
5. Perseverance	-.20***	-.01	.07*	.07*	.65***	.03	.05
6. Intrinsic motivation	-.06	-.09**	-.01	-.03	.04	.72***	.01
7. Goal pursuit	.07*	.11***	.07*	-.03	.07*	.04	.56***
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	0.49	0.33

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

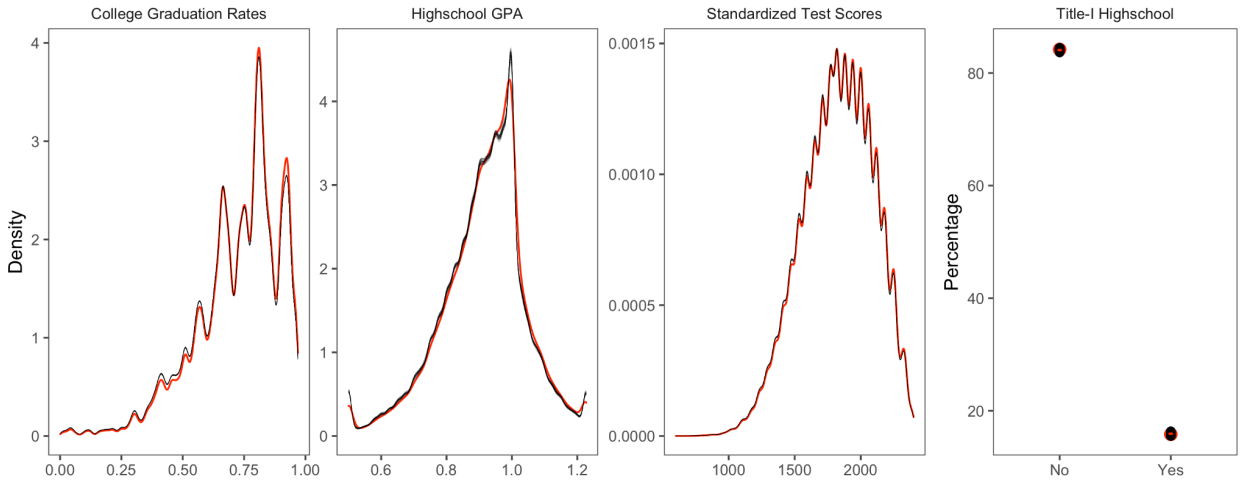
Personal Quality	1	2	3	4	5	6	7
Computer-generated likelihoods							
1. Prosocial purpose	.85***	.00	-.05*	-.10***	-.08**	-.06*	.05
2. Leadership	-.01	.78***	.13***	.00	-.04	-.14***	.04
3. Teamwork	-.06*	.17***	.55***	.07**	.06*	.00	.06*
4. Learning	-.11***	-.02	.04	.76***	.10***	-.02	-.01
5. Perseverance	-.12***	-.03	.04	.09***	.61***	.06*	.02
6. Intrinsic motivation	-.04	-.11***	.00	.01	.06*	.68***	.02
7. Goal pursuit	.07**	.06*	.07**	-.02	.02	.02	.56***
Frequency of human rating	0.34	0.18	0.26	0.43	0.19	0.42	0.31
Mean of computer-generated likelihood	0.36	0.20	0.27	0.47	0.21	0.47	0.34

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

Quality check of imputation 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



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 racerReported		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 highsSchool		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	1.120***	1.071***
	(0.082)	(0.087)
Leadership	1.134***	1.061***
	(0.090)	(0.087)
Teamwork	1.050***	1.004
	(0.082)	(0.083)
Mastery orientation	1.050***	1.034**
	(0.081)	(0.083)
Perseverance	1.080***	1.023
	(0.080)	(0.088)
Intrinsic motivation	1.056***	1.011
	(0.079)	(0.084)
Goal pursuit	1.032**	1.019
	(0.079)	(0.082)
Race/ethnicity (vs. White)		
Black		0.820***
		(0.347)
Latino		0.934
		(0.340)
Asian		0.759***
		(0.300)
Other		0.757***
		(0.305)
No race reported		0.839***
		(0.223)
Parental education (vs. no parent w/ college degree)		
One parent w/ college degree		1.261***
		(0.219)
Two parents w/ college degree		1.454***
		(0.212)
Female		1.383***
		(0.174)
Married parents		1.338***
		(0.196)
English language learner		0.711***
		(0.285)
Title 1 high school		0.910**
		(0.208)
Out-of-school activities		
Number of OSA		1.201***
		(0.085)
Time per OSA		1.082***
		(0.079)
Proportion sports		1.061***
		(0.081)
Standardized test scores		1.241***
		(0.099)
HSGPA		1.441***
		(0.090)
Constant	3.481***	2.449***
	(0.078)	(0.253)
AUC	.555	.702
N	306,463	306,463

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	1.135***	1.063***
	(0.005)	(0.005)
Leadership	1.135***	1.056***
	(0.005)	(0.005)
Teamwork	1.094***	1.033***
	(0.005)	(0.005)
Mastery orientation	1.065***	1.036***
	(0.004)	(0.005)
Perseverance	1.079***	1.000
	(0.005)	(0.005)
Intrinsic motivation	1.061***	1.003
	(0.004)	(0.005)
Goal pursuit	1.025***	0.995
	(0.004)	(0.005)
Race/ethnicity (vs. White)		
Black		0.754***
		(0.020)
Latino		0.857***
		(0.020)
Asian		0.696***
		(0.018)
Other		0.734***
		(0.018)
No race reported		0.828***
		(0.013)
Parental education (vs. no parent w/ college degree)		
One parent w/ college degree		1.156***
		(0.013)
Two parents w/ college degree		1.196***
		(0.012)
Female		1.464***
		(0.010)
Married parents		1.280***
		(0.011)
English language learner		0.683***
		(0.016)
Title 1 high school		0.974*
		(0.014)
Institutional graduation rates		1.890***
		(0.006)
Out-of-school activities		
Number of OSA		1.157***
		(0.005)
Time per OSA		1.082***
		(0.005)
Proportion sports		1.028***
		(0.005)
Standardized test scores		1.164***
		(0.006)
Constant	3.558***	2.989***
	(0.004)	(0.015)
AUC	.562	.741
N	306,463	306,463

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

Appendix A. BERT Settings file

```
# Modify this file as needed
import os
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_selection import SelectKBest
from sklearn.naive_bayes import MultinomialNB
from imblearn.combine import SMOTEENN
from imblearn.over_sampling import RandomOverSampler, ADASYN
from imblearn.under_sampling import RandomUnderSampler
from src.configuration.settings_template import Settings, SettingsEnumOptions
from src.models.nets.windows_of_context_rnn import WindowRnn
from src.pipeline.resampling import DatasetSampler
from src.pipeline.transformers.empty_transformer import EmptyTransformer
from src.pipeline.transformers.standard_scaler_3d import StandardScaler3D
from src.common.meta import MetaUtils
from src.metrics.custom_scorers import PearsonCorrelationScorer

# Folder to output results in, helpful to change if you don't want to overwrite previous results
Settings.IO.RESULTS_OUTPUT_FOLDER = "bert_results_nosj"
Settings.SAVE_MODELS = True
Settings.SKIP_SAVED_MODELS = True

# Names of columns in spreadsheet to identify what should be the input data, and what should be the predicted labels
Settings.COLUMNS.IDENTIFIER = "applicantprofileid"

Settings.COLUMNS.MAKE_ALL_LABELS_BINARY = True
label_list = ["Type_bin", "Accolades_bin", "Connection_bin", "Goal_bin","Goal_r_bin", "Leadership_bin", "Learning_bin", "Persevere_bin",
"Selftrans_bin", "Team_bin"]

Settings.IO.DATA_INPUT_FILE = "common_app_no_sj_Type_bin.csv"
Settings.IO.DATA_INPUT_FILE_PREFIX = "common_app_no_sj_"
Settings.IO.DATA_INPUT_FILE_SUFFIX = ".csv"
Settings.COLUMNS.Y_LABELS_TO_PREDICT = label_list
Settings.FEATURE_INPUT_SOURCES_TO_RUN = [SettingsEnumOptions.LanguageFeatureInput.with_language_from_column("response")]
Settings.BERT_FEATURES.sentence_column_name = "response"
Settings.PREDICTION = SettingsEnumOptions.Prediction.CLASSIFICATION

bert_labels = {
    "multilabel": {
        "is_multilabel": True,
        "num_labels": 10,
        "label_list": label_list,
    },
}

for label in label_list:
    bert_labels[label] = {
        "is_multilabel": False,
        "num_labels": 1,
        "convert_to_onehot": True,
    }

Settings.COLUMNS.BERT_LABELS = bert_labels
Settings.CROSS_VALIDATION.NUM_TRAIN_TEST_FOLDS = 10
Settings.CROSS_VALIDATION.NUM_CV_TRAIN_VAL_FOLDS = 5
Settings.CROSS_VALIDATION.SCORING_FUNCTION = 'roc_auc'
Settings.CROSS_VALIDATION.HYPER_PARAMS.BERT = {
    'epochs': 4,
    'batch_size': 16,
    'max_seq_len': 204
}



Settings.RANDOM_STATE = 42
```

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.