Developing and Evaluating Language-Based Machine Learning Algorithms for Inferring Applicant Personality in Video Interviews

Louis Hickman[1], Rachel Saef[2,] Vincent Ng[3], Sang Eun Woo[1], Louis Tay[1], & Nigel Bosch[4]

[1]Purdue University; [2]Northern Illinois University; [3]University of Houston; [4]University of Illinois at Urbana-Champaign

In press at Human Resource Management Journal for the Special Issue on HR/People Analytics and Human Resource Management

Acknowledgements: The authors thank Kim Soung Hoon, Meg Becker, Simone Silver, Emily Willard, Caroline Mello, Drew Greenagel, Mya Findley, and Amel Chekili for their help with data collection and interviewer-reported personality ratings.

Developing and Evaluating Language-Based Machine Learning Algorithms for Inferring Applicant Personality in Video Interviews

Abstract. Organizations are increasingly relying on people analytics to aid human resources decision-making. One application involves using machine learning to automatically infer applicant characteristics from employment interview responses. However, management research has provided scant validity evidence to guide organizations' decisions about whether and how best to implement these algorithmic approaches. To address this gap, we use closed vocabulary text mining on mock video interviews to train and test machine learning algorithms for predicting interviewee's self-reported (*automatic personality recognition*) and interviewer-rated personality traits (*automatic personality perception*). We use 10-fold cross-validation to test the algorithms' accuracy for predicting Big Five personality traits across both rating sources. The cross-validated accuracy for predicting self-reports was lower than large-scale investigations using language in social media posts as predictors. The cross-validated accuracy for predicting interviewer ratings of personality was more than double that found for predicting self-reports. We discuss implications for future research and practice.

Keywords: text mining, machine learning, big five, personality traits, video interviews, LIWC**,** cross-validation, elastic net regression

Practitioner notes.

What is currently known

- People analytics tools are being marketed to organizations that purport to automatically infer interviewee characteristics.
- Available evidence suggests self-reported personality can be inferred from social media language.
- Yet, the validity of such approaches for applicant screening in video interviews is unknown.

What this paper adds

- We developed algorithms on video interviews to infer interviewee personality from their computer transcribed interview responses.
- We inferred both self-reported and interviewer-rated personality.
- Interviewer-rated personality can be inferred with much greater accuracy than self-reported personality.

The implications for practitioners

- Algorithmic approaches for scoring interviewee attributes may save organizations time and money.
- Algorithms trained on interview data may function better than off-the-shelf algorithms, and investigating how algorithms were built and designed is important for legal defensibility.
- More validity evidence is needed before algorithmic personality inference should be adopted by organizations.

Developing and Evaluating Language-Based Machine Learning Algorithms for Inferring Applicant Personality in Video Interviews

Organizations are increasingly relying on people analytics to improve human resources (HR) decision-making. People analytics applies advanced statistical and computational methods to organizational data to inform HR decisions. For instance, organizations are increasingly relying on machine learning (ML) algorithms within selection contexts with hopes of improving efficiency and reducing the influence of human bias (Oswald et al., 2020). Such algorithms automatically infer applicant knowledge, skills, abilities, and other characteristics (KSAOs; Angrave et al., 2016), such as personality traits (Rotolo et al., 2018).

Traditionally, organizations have relied on self-reports to assess personality in selection. However, concerns about socially desirable responding and faking (Vazire, 2010) have led to calls for alternatives to self-reports for assessing personality (e.g., Morgeson et al., 2007; Ployhart et al., 2017). Interviewer personality judgments are one such alternative and may better predict job performance than self-reports (Levashina et al., 2014; Van Iddekinge et al., 2005). Yet, personality assessment usually occurs early in the screening process, making it prohibitively costly to replace self-reports with interviewer trait assessments. Thus, using people analytics to automate interviewer-based personality assessments on a large scale holds potential to improve the utility of hiring outcomes (Chamorro-Premuzic et al., 2017).

Although ML holds promise for efficiently and accurately inferring applicant KSAOs, scant empirical evidence exists to support the validity of algorithmic approaches for personnel assessment. Human resources science must investigate the validity of algorithmic approaches to guide organizations' decisions about whether and how best to implement them in practice. Researchers in other fields, like computer science, actively research ML-based assessments (Rotolo et al., 2018) and may benefit from HR's extensive expertise in psychometrics and scale development. Therefore, the current study examines the validity of using interviewee responses to selection interview questions (i.e., verbal behavior) as predictors in ML algorithms to automatically score applicants' Big Five personality traits: extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience.

Outside of HR literature, researchers have applied ML to develop language-based models for automatically inferring personality traits from language use (i.e., automatic personality recognition; e.g., IBM, 2019; Schwartz et al., 2013). These approaches are based on the idea that personality affects both *what* people talk about and *how* they talk about it, which results in between-person differences in patterns of language use (Tausczik & Pennebaker, 2010). While these models could be applied to interviews, the existing language-based models for inferring personality were developed using social media language. Such models show little evidence of accurate personality recognition or perception in interview contexts (Hickman et al., 2019). This aligns with previous findings that personality measures better predict outcomes when they are contextualized to similar contexts as the outcome (Shaffer & Postlethwaite, 2012). Therefore, ML algorithms should be developed based on selection interview responses to contextualize these assessments to work contexts. Interviewee responses (i.e., verbal behavior) form the key behavioral component of interviews and are the primary source of information for interviewer ratings, particularly in structured interviews that use behaviorally anchored rating scales.

To our knowledge, existing research using natural language and digital footprints to predict personality traits has only developed models for self-reported personality (Azucar et al., 2018). However, researchers have suggested that interviewer-rated personality traits may better predict job performance than self-reported traits (Levashina et al., 2014). Personality has two different components: 1) the relatively enduring patterns of internal feelings, thoughts, and behaviors (Roberts & Jackson, 2008) that reflect a person's inner nature (i.e., identity); and 2) a person's social reputation, or the way one is perceived by others (i.e., reputation; Hogan, 1991). Although concerns exist regarding the impact of self-presentation on how interviewers perceive interviewee personality (e.g., impression management; Van Iddekinge et al., 2005), self-reports are also distorted by self-presentation, even in the absence of motivation to fake (Hogan et al., 1996). As interviewees' self-reported and interviewer-rated personality ratings each provide valuable, unique information for predicting future behavior (e.g., job and academic performance; Connelly & Chang, 2016), focusing on self-reports does not represent the full relationship between language use and personality traits. Therefore, the current study applies closed vocabulary text mining to mock interviews, then trains and tests the accuracy of interview-native language-based algorithms for automatic personality recognition (i.e., inferring interviewee self-reported traits) and automatic personality perception (i.e., inferring interviewer-rated traits; Vinciarelli & Mohammadi, 2014).

This study contributes to the selection and management literatures in several ways. First, this study answers calls to investigate alternatives to self-report personality measures (Morgeson et al., 2007) and technologies for automatically scoring applicant KSAOs (Chamorro-Premuzic et al., 2017). The scientific study of such technologies can ensure HR and management *science* keeps pace with and remains relevant to management *practice*. Second, we integrate insights from two related research streams: 1) the use of ML to score personality traits from digital footprints (e.g., Azucar et al., 2018), and 2) the application of text mining and ML to automate existing selection and assessment procedures (e.g., Campion et al., 2016; Speer, 2018). To our knowledge, the present investigation is one of the first studies to examine the predictive accuracy of language-based interview-native algorithms for predicting interviewee personality, and it expands the body of work predicting personality from language use by engaging in both automatic personality recognition and perception. Third, the current paper discusses the conceptual concerns and practical benefits of using such data-driven approaches for inferring applicant personality. Such approaches present attractive potential benefits in terms of time and cost savings, yet we urge caution in their adoption until further utility and validity evidence is available. Future work is crucial for ensuring that inferences about applicant personality and hiring decisions based on these methods reduce (rather than exacerbate) biases compared to traditional selection procedures.

Language Use and Personality

The current study trains and tests algorithms to utilize language use in a video interview to predict personality traits. Like other behaviors, language use is a function of both personality and the situation (Mischel & Shoda, 1995). Therefore, within a given context, personality traits should relate to language use, and this is what researchers have found in a variety of contexts, including everyday conversations (Mehl et al., 2006), personal essays (Pennebaker & King, 1999), blogs (Yarkoni, 2010), and social media posts (Schwartz et al., 2013). Additionally, management researchers have theorized that individual differences, including personality, directly affect interviewee responses regardless of their qualifications, experience, or other job-relevant KSAOs (Huffcutt et al., 2011).

Further, personality traits affect both *what* people talk about and *how* they talk about it (Tausczik & Pennebaker, 2010). *Content words*, or what people talk about, tend to vary across contexts and include nouns, verbs, adjectives, and adverbs. The style of speech, or how people talk, tends to be more stable across contexts. The style of speech is conveyed primarily through *function words*, including articles (e.g., a, the), auxiliary verbs (e.g., am, will), conjunctions (e.g., and, but), prepositions (e.g., to, with), and pronouns (e.g., I, she). Function words comprise only .05% of all words in the English language, but they make up over half the words used in our speech and writing (Tausczik & Pennebaker, 2010).

Language use can be analyzed in various ways (Hickman et al., 2020). For the present study, we adopted closed vocabulary text mining. In closed vocabulary text mining, word lists are compiled in dictionaries that reflect meaningful psychological categories. Those words are counted to score the proportion of text corresponding to each category (McKenny et al., 2018). In the present study, we adopted the well-known closed vocabulary tool, Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015). Besides its popularity, LIWC is part of a tool that assesses leaders and tracks organizational culture from language use in emails, speeches, and press releases (Receptiviti, n.d.). Additionally, numerous researchers have analyzed how language, as operationalized by LIWC, relates to personality traits. Table 1 summarizes the significant relationships between the Big Five traits and LIWC categories that have been observed in multiple studies. Some previously observed relationships make clear conceptual sense, including extraverts talking more about social processes and emotionally stable people talking less about negative emotions and anxiety (Yarkoni, 2010). However, some relationships changed direction across studies, suggesting that the relationship between traits and language use is context-specific, a function of both the person and the situation (Mischel & Shoda, 1995). Therefore, the relationships between language use and traits, which form the basis for automatic personality recognition and perception algorithms, are likely context-bound.

Method

Sample and Procedure

We collected mock video interviews from 490 undergraduate psychology students (246 female) at a large university in the Midwestern United States. The students averaged 19 years old (SD = 18.85) and had previously interviewed for actual positions 2.68 times (1.90 in-person interviews, 0.54 phone interviews, and 0.24 video interviews).

Participants completed an online survey and self-reported their Big Five traits. Then, to gain interviewing experience, they participated in a one-way mock interview consisting of three interview prompts (*Please tell us about yourself*; *Please tell us about a time you demonstrated leadership*; and *Please tell us about a time you worked effectively in a team*). Participants were encouraged to take time to prepare, then recorded their response to each prompt. Three subject matter experts well-versed in personality and personnel selection designed the prompts to be broad and applicable to various professions. Participants were instructed to answer each prompt for a minimum of two and a maximum of three minutes, for a total interview length of six to nine minutes (M = 6 min 51 s; word count M = 951.21). Four hundred sixty-seven participants completed the study in full. Twenty-six videos could either not be transcribed or viewed due to technical difficulties experienced during the study, resulting in a final sample of 441 participants.

Self-reported personality

Participants self-reported personality using Goldberg’s (1992) 50-item measure of the Big Five Factor Markers (BFFM), available in the International Personality Item Pool (Goldberg, 1999). We dropped 52 participants’ self-reports for failing attention checks, leaving 389 self-reports. Cronbach’s alpha for self-reported traits ranged from .76 for openness to .90 for extraversion, as reported in Table 2.

Interviewer-rated personality

Undergraduate research assistants watched the mock video interviews and provided ‘interviewer’ ratings of interviewee traits using an observer version of the Ten Item Personality Inventory[1] (Gosling et al., 2003). Before doing so, research assistants participated in two hours of frame of reference training. It included defining the Big Five traits, explaining the scale and scale anchors, watching mock video interviews, assigning practice ratings, and discussing specific, observed behaviors that lead to (dis)agreement in ratings. Research assistants were instructed not to rate participants if they were previously acquainted. At least three research assistants from a pool of eight watched and rated each interviewee. We chose to watch and rate the video interviews of participants who failed attention checks but discarded their self-reports, resulting in interviewer-rated personality for 441 interviewees. We averaged all available interviewer ratings before analysis. ICC(C, 8) ranged from .66 (emotional stability) to .89 (extraversion), as reported in Table 2.

Language data

We transcribed participants’ full mock video interview responses using IBM Watson Speech-to-Text (IBM, 2019). Although computerized transcription can introduce errors, we thought it essential to use computerized transcriptions because similar products sold to organizations use automatic, computerized transcriptions (Kutik, 2015). We analyzed interviewee transcriptions using 75 directly counted categories from LIWC (Pennebaker et al., 2015). LIWC counts words from conceptually derived categories and scores them as the proportion of the overall response. Therefore, scores for LIWC categories (except word count)

[1]While it is not ideal to use different scales for self- and interviewer-reports of personality, the BFFM and TIPI provide comparable assessments of the Big Five (Donnellan et al., 2006) and show similar patterns of relationships with workplace behavior (Burns et al., 2017). Further, the TIPI is based on the BFFM and converges with the BFFM measures in self-, observer-, and peer-reports (Gosling et al., 2003). Pragmatically, the time required for interviewers to rate traits increases considerably with longer instruments, and personality judgments in employment interviews are often based on one item per interview question (e.g., Van Iddekinge et al., 2005).

indicate the proportion of words spoken across the entire interview that fell into each predefined category. We did not include LIWC punctuation variables, as the text was spoken, not written. Additionally, we did not include LIWC categories with very low base rates and, therefore, low variability, including *swear words, fillers* (e.g., I mean, you know), *netspeak* (e.g., btw, lol), and *death* (e.g., bury, coffin).

Machine learning prediction

Predictive modeling. We entered LIWC category scores as predictor variables and personality trait ratings as outcome variables in our models. We trained and tested 10 separate ML models using the *caret* R package (Kuhn, 2008): one for predicting each of the self-reported Big Five traits, and one for predicting each of the interviewer-rated Big Five traits. For all ten models, we adopted elastic net regression and 10-fold cross-validation (described below) to train and test the predictive accuracy of LIWC categories (i.e., language-based personality inference). Psychologists have referred to this process as statistical learning (Chapman et al., 2016), wherein regression-based algorithms with numerous potential predictors are tuned to maximize cross-validated accuracy for a given outcome. This step is an inductive, data-driven approach that serves as a starting point for measurement refinement and theory development (Jebb et al., 2017). Data-driven approaches allow exploring all potential interviewee language (e.g., Park et al., 2015) rather than limiting automatic personality recognition and perception to traditional conceptualizations of personality. Given the nascent understanding of how language use is associated with the Big Five traits in evaluative contexts, it is important to consider all LIWC categories as predictors.

Elastic net regression. Elastic net regression was chosen for the predictive algorithm because it has two regularization terms that shrink coefficients towards zero to prevent overfitting (Chapman et al., 2016; Zou & Hastie, 2005). The regularization terms address the bias-variance tradeoff: they are tuned by varying the two hyperparameters (alpha and lambda) to determine each regularization term's optimal weight, resulting in cross-validated accuracy that performs favorably compared to other algorithms for personnel assessment purposes (Putka et al., 2018). By varying alpha, elastic net regression can act as a) ridge regression, b) least absolute squares shrinkage and selection operator (LASSO) regression, or c) a hybrid of the two. Ridge regression forces coefficients *toward* zero to reduce prediction variance and, therefore, error. LASSO regression forces coefficients *to* zero in response to predictor multicollinearity and model complexity, thereby removing some predictors from the model. When alpha equals zero, elastic net is ridge regression, and when alpha equals one, elastic net is LASSO regression. When alpha is greater than zero but less than one, elastic net acts as a hybrid of the two models, both shrinking coefficients toward zero and forcing some to zero. Lambda determines the severity of regression weight shrinkage, such that larger values result in greater shrinkage. Therefore, higher alpha and lambda values increase regression coefficient regularization to reduce model complexity and overfitting to increase cross-validated accuracy. To train the predictive models, we systematically varied alpha and lambda (we tried 10 values of each in all models, using default values from caret). Then, we selected the final model based on which combination of values provided the highest average cross-validated correlation between predicted and reported traits. We used correlations for hyperparameter tuning instead of error rates (e.g., mean absolute error) because correlations are scale-independent and more familiar to management scholars.

Cross-validation strategy. We adopted 10-fold cross-validation, a form of *k*-fold cross-validation where *k*=10, to estimate algorithm accuracy for each of the ten predictive models. *k*-fold cross-validation involves splitting the data into *k* folds, training predictive models on *k*-1 folds (the *training* dataset), then testing the accuracy of the model's predictions on the remaining fold (the *testing* dataset). This process is repeated *k* times, with each fold used only once for testing. By splitting the data into *k* folds, *k*-fold cross-validation mitigates the impact of sampling error on accuracy estimates by using all data (rather than only a subset of data) to test predictive accuracy. We chose k=10 following recent recommendations (Bleidorn & Hopwood, 2019). When sample size exceeds 300, 10-fold cross-validation provides reliable estimates of model generalizability (Putka et al., 2018). Estimating accuracy in 10-fold cross-validation involves calculating the average correlation between predicted and reported traits across the 10 test folds for each set of hyperparameters, then reporting these correlations for the optimal hyperparameters. These tests provide management scholars and practitioners with an initial estimate of the potential accuracy of language-based personality inference.

Results

The descriptive statistics of participant gender, self- and interviewer-rated personality, and criteria are presented in Table 2. The average convergence between interviewee self-reported and interviewer-rated personality was M_r = .24, slightly smaller in magnitude but not significantly different than the convergence found between self- and interviewer ratings in 30-minute long face-to-face mock interviews (e.g., M_r = .28; z = .33; p = .74; Barrick et al., 2000). The correlation between self- and interviewer-rated conscientiousness was not significant (r = .06, p = .26).

Heteromethod-monotrait convergence (i.e., same trait correlations between self- and interviewer-ratings) is frequently used as a metric of personality perception accuracy, but it is suboptimal because self-reports and other-reports reflect different information (i.e., identity vs. reputation; Hogan, 1991). To further investigate the accuracy of self- and interviewer-rated traits, we conducted a series of hierarchical regressions predicting academic criteria (i.e., self-reported high school grade point average, SAT verbal, SAT math, and ACT scores). In the first step, we controlled for gender and added either the five self-reported or interviewer-rated traits. Then, in the second step, we added the other five trait estimates (i.e., when interviewee self-reports were added in step one, interviewer ratings were added in step two, and vice versa). Full results are provided in Appendix A. Across the four outcomes, self- and interviewer-rated traits significantly increased R^2 for three of the outcomes beyond the other personality rating source. On average, interviewer ratings explained more variance in these outcomes than did self-reports. Taken together, these two pieces of evidence regarding interviewer-rated traits support the idea that the mock interviews provided personality relevant information, and interviewers provided accurate personality judgments.

Before summarizing our ML investigation results, we present the significant correlations between LIWC categories and both self- and interviewer-rated traits in Table 3. Many of the significant correlations align with prior research summarized in Table 1. More significant correlations were observed between LIWC categories and interviewer-rated personality (average number of significant correlations M_{self} = 9; $M_{interviewer}$ = 21.8). Among categories that were significantly related to both self- and interviewer-ratings for a given trait, the sign of the relationship flipped twice: *leisure* was positively related to conscientiousness self-reports but negatively to interviewer ratings, and *informal* was positively related to emotional stability self-reports but negatively to interviewer ratings. Bivariate correlations are presented instead of predictor regression weights due to algorithmic uncertainty, or uncertainty due to error in estimating personality traits from text data (Kennedy & O'Hagan, 2001). Specifically, the unique weights and rankings of LIWC predictors can shift across the 10-fold cross-validated models due to sampling error associated with using different data to train each model. Additionally, each of the 10-fold cross-validated models for a single trait does not necessarily include the same set of LIWC predictors. Therefore, bivariate correlations are more appropriate for examining these relationships, as they use the full information available in the sample and can be compared to prior findings.

We now evaluate the predictive algorithms. Table 4 reports: the optimal hyperparameters; the average convergent correlation between each models' predictions and reported personality across the 10 test folds[2]; the minimum, maximum, and standard deviation of the 10 convergent correlations; and the average convergent correlation corrected for unreliability. The upper portion of Table 4 reports this information for self-reports, and the lower portion reports this information for interviewer ratings. Across the Big Five, the average convergence between language-based predictions and self-reported traits was $\bar{r}$ = .19 ($\bar{\rho}$ = .20 correcting for self-report unreliability). The highest accuracy was observed for extraversion (*r* = .27), and the lowest accuracy was observed for openness to experience (*r* = .12).

For interviewer-reports, the average convergence across the Big Five traits between language-based predictions and interviewer-rated traits was $\bar{r}$ = .39 ($\bar{\rho}$ = .45 correcting for interrater unreliability). The highest accuracy was observed for extraversion (*r* = .49), and the lowest accuracy was observed for emotional stability (*r* = .21).

Following a reviewer's suggestion, we examined whether convergence differed for male and female interviewees. To do so, we calculated the correlation between predicted and observed traits for males and females separately, converted the correlations to Fisher's *z*-score, then compared them using a two-tailed test. In only one of the ten cases was the difference marginally

[2] We compared these results to random forest. The results were nearly identical, as the average convergence of random forest predictions was $\bar{r}$ = .17 for self-reports and $\bar{r}$ = .38 for interviewer-reports.

significant: for interviewer-reports of conscientiousness, the ML models were more accurate at predicting judgments of males' than females' conscientiousness (r_{men} = .46; r_{women} = .30; p = .05).

Discussion

Organizations are increasingly applying algorithmic assessments to employment interviews to automatically score applicant KSAOs from interviewee responses (i.e., verbal behavior). However, the off-the-shelf, commercially available language-based algorithms for automatically scoring personality are often developed on social media datasets (Tay et al., 2020). Recent research found that social media-based language models do not accurately assess personality in the interview context (Hickman et al., 2019). The current study applied closed vocabulary text mining and ML to examine whether language-based algorithms trained on interviewee verbal behavior could accurately infer interviewee personality traits, both self- and interviewer-rated. Below we discuss a few major findings and their implications.

First of all, our results showed that the accuracy of algorithms that used interviewee language to predict self-reported personality ($\bar{r}$ = .19) was lower than the average convergence between self- and interviewer-rated Big Five traits in the present study ($\bar{r}$ = .24). Additionally, the average accuracy was .07 lower than Schwartz et al.'s (2013) application of LIWC to predict personality in a sample of more than 70,000 Facebook users. Unique from previous research, we also developed language-based models to predict interviewer-rated personality from a mock video interview. The language-based algorithms were, on average, twice as accurate at predicting interviewer ratings ($\bar{r}$ = .39) than predicting self-reports, an important consideration since interviewer-rated personality may better predict job performance (Levashina et al., 2014).

Second, for both self- and interviewer-rated personality, LIWC variables were more strongly related to more observable traits. This pattern converges with previous findings that highly observable (e.g., extraversion) personality traits demonstrate higher interrater reliability and self-observer convergence than less observable traits (e.g., neuroticism, openness; Connelly & Ones, 2010). Given that reliability puts a ceiling on validity, the lower validity of language use for less observable traits makes sense. The Self-Other Knowledge Asymmetry (SOKA) model (Vazire, 2010) helps explain why. The SOKA model posits that less observable traits that characterize internal cognitive processes and affective tendencies (e.g., neuroticism) are more accurately judged by the self because individuals have unique information into their own thoughts and feelings that are not accessible to others. Conversely, the SOKA model posits that evaluative traits (e.g., agreeableness, conscientiousness) are more accurately judged by observers because of self-bias that motivates distortions in self-reports. Both the self and others have information about visible, non-evaluative traits like extraversion that are expressed in more behavioral (e.g., talkative) ways. In the present study, the strongest correlation between any Big Five personality trait score and LIWC predictor was between interviewer-reported extraversion and word count (i.e., number of words used by interviewee, r = .45; Table 3).

Third, the LIWC scores, based on predefined categories of *interviewee* word usage, showed stronger and more numerous relationships with *interviewer ratings* than self-reports for all traits. Therefore, interviewee language use appears to play a larger role in how observers (e.g., interviewer) form perceptions of target (e.g., interviewee) personality than they are manifestations of target personality. This is important because although people think that self-reports will be more predictive of behavior, self- and other-rated personality traits are, on average, approximately equally predictive of behavior, although they differ in which behaviors they are most predictive (Vazire & Mehl, 2008). Thus, both operationalizations of personality provide overlapping but substantively unique information.

Additionally, meta-analytic findings indicate that other-reported Big Five traits can be more predictive of academic and job performance than self-reported traits (Connelly & Ones, 2010). However, it is worth noting that the other-rated trait best predicted by interviewee language use (extraversion) is not consistently related to job performance relative to conscientiousness (Connelly & Ones, 2010). Finally, meta-analytic results have shown that self-report common method variance *attenuates* the ability of the traits to predict both academic and job performance due to response style distortion (Connelly & Chang, 2016). Thus, since interviewee language (operationalized as LIWC variables) and academic criteria are more strongly correlated with interviewer-rated traits, the interviewer ratings may be capturing more accurate personality trait scores. An alternative (although not mutually exclusive) explanation for the higher accuracy of automatic personality perception compared to automatic personality recognition is that interviewer ratings and algorithmic scores share a common information source: the interview responses. Because interviewer ratings of personality were made based on the interview responses, whereas self-reports asked respondents to rate their personality in general, algorithmic scores based on the LIWC variables may be predicting the former more strongly because both are tapping into contextualized rather than general personality.

Taken together, our findings suggest that automatic, language-based personality inference in employment interviews holds potential for complementing traditional self-reports and interviewer-ratings by reducing the time and cost associated with early-stage applicant screening. However, we recommend that organizational decision-makers carefully weigh the convenience of automatic inference against the imperfect prediction it provides: These algorithms attempt to replicate human-rated personality, but their predictions correlate only moderately with human raters. As a result, the algorithms may provide less valid predictions of workplace outcomes while possibly exacerbating preexisting biases in the human ratings. There are many other practical challenges to implementing and monitoring these ML-based assessments that organizations must carefully consider. For example, issues of faking in algorithmic vs. traditional interview assessments have not received systematic investigations. Before algorithmic interviews can be adopted, therefore, it is necessary to do a thorough cost-benefit analysis comparing algorithmic and traditional assessments on multiple aspects (SIOP, 2018).

Limitations and Future Research

While this study has several strengths (e.g., parallel testing of automated personality recognition and perception), there are also notable limitations. First, the use of a non-applicant sample limits the conclusions we can draw. Compared to mock interviews, employment interviews with real applicants typically function in contexts with higher stakes, which represents a stronger situation (Meyer et al., 2010). Employment interviews have significant consequences for individuals, whereas mock interviews do not. Such strong situations limit the influence of individual differences on behavior because interviewees are motivated to engage in self-presentation regardless of predisposition (Van Iddekinge et al., 2007). Therefore, self-reported personality may be even less related to interviewee behavior in applicant samples, and future research should investigate whether similar accuracy can be obtained in applicant samples.

Second, employment interviews tend to be much longer than those in the present study. Indeed, one tool being marketed to organizations that purports to automatically infer interviewee KSAOs appears to require an average interview length of 15-20 minutes (Mondragon et al., 2019). Relatedly, prior investigations of language-based personality inference have sometimes excluded participants with fewer than 1,000 available words (e.g., Schwartz et al., 2013). Therefore, the relatively short length of the interviews investigated here (on average, 6 min 51 s and 951 words) likely attenuated the accuracy of the approach. Longer interviews may increase the accuracy of language-based personality algorithms. Another point of concern with our current study design is the use of student raters and somewhat lower reliabilities for interviewer-rated personality, which can substantially attenuate correlations (Connelly & Ones, 2010). Therefore, it may be beneficial for future research to include experienced interviewers, longer scales, and more raters for training the algorithms to maximize the potential that automatically inferred traits have for predicting workplace criteria.

At the same time, it is also important to note that even with longer interviews and trained raters (e.g., professional recruiters), self- and interviewer-ratings may not perfectly converge due to the unique perspectives that interviewee and interviewer bring (Connelly & Ones, 2010; Funder & West, 1993; Vazire, 2010). Indeed, prior studies with such conditions found non-significant convergence between conscientiousness self-reports and interviewer ratings (Barrick et al., 2000). In our data, despite low convergence, interviewer-rated conscientiousness had acceptable interrater reliability (Table 2) and predicted academic criteria (Appendix A, Table A2), meaning that these ratings captured substantive and useful information about personality traits.

Third, our study focused on the convergence of our ML-based assessments with self- and interviewer-ratings of the same construct. Yet convergent relationships represent just one piece of evidence that can be used to make judgments about a measure's validity (Bleidorn & Hopwood, 2019; SIOP, 2018). Specifically, the extent to which such approaches exhibit discriminant evidence (including factorial validity among the ML trait estimates) and predict workplace criteria will be necessary for understanding whether such approaches can be validly applied to personnel selection. Given the rationale for this paper, automated personality scores' relationships with job performance and other organizational criteria will be particularly critical for not only the theoretical understanding of score meaning but also in pragmatically justifying their use in personnel selection.

Beyond addressing these key limitations, we also suggest a few additional research directions that may stem from the current study. First, it is crucial that ML-based screening methods are fair and equally accurate across demographic groups (SIOP, 2018). We tested if the convergent-related validity for the assessments developed in the present study differed for men and women. For interviewer-reported conscientiousness, the ML models were somewhat more accurate judging men than women. Concerns have been raised that algorithmic interview assessments may discriminate against other legally protected groups (Harris et al., 2018). Therefore, future work should also test for other types of demographic (e.g., race, gender) bias and investigate strategies for reducing such biases.

Second, while our study used computerized transcription instead of manual transcription to enhance ecological validity (Kutik, 2015), computerized transcription is not error-free. Errors introduced by the transcription software may reduce validity. Additionally, computerized transcription may have higher error rates for some demographic groups (e.g., interviewees whose first language is not English). Future research should seek to better understand the error rate of computerized transcriptions, and the effect on accuracy, by directly comparing how manual and computerized transcription influences the relationship between language use and personality traits. In addition, given that computerized transcription errors can inaccurately record words due to low speech clarity and volume, doing so could test if these speech differences also influence interviewer ratings and if this negatively affects automatic scores for particular groups (e.g., non-native English speakers).

Third, as mentioned above, closed vocabulary text mining is just one way of analyzing natural language data (Hickman et al., 2020). Closed vocabulary text mining counts conceptually related words to score psychologically meaningful categories. On the other hand, open vocabulary text mining counts words and phrases with no preformed notions about how they relate to each other or to outcomes, allowing all words and phrases to be used as predictors (Kern et al., 2014). Compared to open vocabulary, closed vocabulary text mining is more precise and easier to summarize than open vocabulary text mining, but it tends to have lower predictive accuracy (e.g., Schwartz et al., 2013). The predictive accuracy of interviewee personality inference may be improved by developing dictionaries specifically designed to tap interview-relevant language. Alternatively, open vocabulary text mining may provide higher accuracy due to its greater flexibility.

Finally, given that only a few of the Big Five personality traits have proven to predict job performance across occupations (e.g., Connelly & Ones, 2010), future research should investigate the validity of ML for assessing other KSAOs. Specifically, based on their robust relationships with important workplace outcomes (e.g., job performance, turnover, counterproductivity; Lievens & Sackett, 2012; Salgado & Moscoso, 2019; Van Iddekinge et al., 2011), we suggest future research apply ML to capture cognitive ability, interpersonal skills, and vocational interests. Additionally, given that interviewers watched video recordings that included facial expressions (i.e., nonverbal behavior) and audio that captured *how* answers were delivered (i.e., paraverbal behavior), future work should also examine potential nonverbal and paraverbal behaviors associated with these individual differences. Huffcutt et al.’s (2011) model of interview performance positions all three types of interviewee behavior (i.e., verbal, nonverbal, and paraverbal) as mediators between interviewee attributes and interviewer ratings. Including all three types of behavior simultaneously as predictors may provide more accurate automated personality inferences.

References

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M., & Stuart, M. (2016). HR and analytics: Why HR is set to fail the big data challenge. *Human Resource Management Journal*, *26*, 1-11. https://doi.org/10.1111/1748-8583.12090

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*(September 2017), 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Barrick, M. R., Patton, G. K., & Haugland, S. N. (2000). Accuracy of interviewer judgments of job applicant personality traits. *Personnel Psychology*, *53*(7), 925–951. https://doi.org/10.1111/j.1744-6570.2000.tb02424.x

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190–203. https://doi.org/10.1177/1088868318772990

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975. https://doi.org/10.1037/apl0000108

Chamorro-Premuzic, T., Akhtar, R., Winsborough, D., & Sherman, R. A. (2017). The datafication of talent: how technology is advancing the science of human potential at work. *Current Opinion in Behavioral Sciences*, *18*, 13–16. https://doi.org/10.1016/j.cobeha.2017.04.007

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods*, *21*(4), 603–620. https://doi.org/10.1037/met0000088

Connelly, B. S., & Chang, L. (2016). A meta-analytic multitrait multirater separation of substance and style in social desirability scales. *Journal of Personality*, *84*(3), 319–334. https://doi.org/10.1111/jopy.12161

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6), 1092–1122. https://doi.org/10.1037/a0021212

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality*, *61*(4), 457–476. https://doi.org/10.1111/j.1467-6494.1993.tb00778.x

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.

Gosling, S. D., Rentfrow, P. J., & Swann Jr., W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Harris, K. D., Murray, P., & Warren, E. (2018). *Letter to U.S. Equal Employment OpportunityCommission regarding risks of facial recognition technology*. Retrieved from https://www.scribd.com/document/388920670/SenHarris-EEOC-Facial-Recognition-2

Hickman, L., Tay, L., & Woo, S. E. (2019). Validity investigation of off-the-shelf language-based personality assessment using video interviews: Convergent and discriminant relationships with self and observer ratings. *Personnel Assessment and Decisions*, *5*(3), 12–20.

Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2020). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods*, 1-33. https://doi.org/10.1177/1094428120971683

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist*, *51*(5), 469–477. https://doi.org/10.1037//0003-066X.51.5.469

Hogan, R. T. (1991). Personality and personality measurement. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology (Vol. 2)* (2nd ed., pp. 873–919). Palo Alto, CA: Consulting Psychologists Press, Inc.

Huffcutt, A. I., Van Iddekinge, C. H., & Roth, P. L. (2011). Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review*, *21*(4), 353–367. https://doi.org/10.1016/j.hrmr.2011.05.003

Jebb, A. T., Parrigon, S., & Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, *27*(2), 265-276.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*(4), 521–551.

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Dziurzynski, L., Ungar, L. H., Stillwell, D. J., … Seligman, M. E. P. (2014). The online social self: An open vocabulary approach to personality. *Assessment*, *21*(2), 158–169. https://doi.org/10.1177/1073191113514104

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.1053/j.sodo.2009.03.002

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. a. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241–293. https://doi.org/10.1111/peps.12052

Lievens, F., & Sackett, P. R. (2012). The validity of interpersonal skills assessment via situational judgment tests for predicting academic success and job performance. *Journal of*

Applied Psychology, *97*(2), 460–468. https://doi.org/10.1037/a0025741

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, *27*(2), 277–290. https://doi.org/10.1016/j.hrmr.2016.08.005

McKenny, A. F., Aguinis, H., Short, J. C., & Anglin, A. H. (2018). What doesn't get measured does exist: Improving the accuracy of Computer-Aided Text Analysis. *Journal of Management*, *44*(7), 2909–2933. https://doi.org/10.1177/0149206316657594

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*(5), 862–877. https://doi.org/10.1037/0022-3514.90.5.862

Meyer, R. D., Dalal, R. S., & Hermida, R. (2010). A review and synthesis of situational strength in the organizational sciences. *Journal of Management*, *36*(1), 121–140. https://doi.org/10.1177/0149206309349309

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*(2), 246–268. https://doi.org/10.1037/0033-295X.102.2.246

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1744-6570.2007.00089.x/full

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in Industrial-Organizational Psychology and Human Resource Management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*(1). https://doi.org/10.1146/annurev-orgpsych-032117-104553

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. P. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, *108*(6), 934–952. https://doi.org/10.1037/pspp0000020

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin: University of Texas at Austin.

Pennebaker, J. W., & King, L. a. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 Years of selection and recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, *102*(3), 291–304. https://doi.org/10.1037/apl0000081

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Roberts, B. W., & Jackson, J. J. (2008). Sociogenomic personality psychology. *Journal of Personality*, *76*(6), 1523–1544. https://doi.org/10.1111/j.1467-6494.2008.00530.x

Rotolo, C. T., Church, A. H., Adler, S., Smither, J. W., Colquitt, A. L., Shull, A. C., … Foster, G. (2018). Putting an end to bad talent management: A call to action for the field of industrial and organizational psychology. *Industrial and Organizational Psychology*, *11*(2), 176–219. https://doi.org/10.1017/iop.2018.6

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for five performance criteria: Hunter and Hunter (1984) revisited. *Frontiers in pPsychology*, *10*, 2227.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, *65*(3), 445–493. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333. https://doi.org/10.1111/peps.12263

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and validity issues in machine learning approaches to personality assessment: A focus on social media text mining. *European Journal of Personality*, *34*(5), 826-844.

Van Iddekinge, C. H., McFarland, L. A., & Raymark, P. H. (2007). Antecedents of impression management use and effectiveness in a structured interview. *Journal of Management*, *33*(5), 752–773. https://doi.org/10.1177/0149206307305563

Van Iddekinge, C. H., Raymark, P. H., & Roth, P. L. (2005). Assessing personality with a structured employment interview: Construct-related validity and susceptibility to response inflation. *Journal of Applied Psychology*, *90*(3), 536–552. https://doi.org/10.1037/0021-9010.90.3.536

Van Iddekinge, C. H., Roth, P. L., Putka, D. J., & Lanivich, S. E. (2011). Are you interested? A meta-analysis of relations between vocational interests and employee performance and turnover. *Journal of Applied Psychology*, *96*(6), 1167-1194.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*(2), 281–300. https://doi.org/10.1037/a0017908

Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, *95*(5), 1202–1216. https://doi.org/10.1037/a0013314

Vinciarelli, A., & Mohammadi, G. (2014). A survey of personality computing. *IEEE Transactions on Affective Computing*, *5*(3), 273–291. https://doi.org/10.1109/TAFFC.2014.2330816

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x

Table 1

Relationships in prior studies between FFM traits and LIWC variables

	LIWC variables		
	Positively correlated	Negatively correlated	Conflicting findings
Extraversion	•Social processes •Family •Friends •Sexual •Affect •Positive emotions •Inclusive •First person singular pronouns	•Tentative •Negations •Articles •Impersonal pronouns	•Words > six letters •Numbers •Work •Perceptual processes
Agreeableness	•Family •Inclusive •Positive emotions	•Negations •Swear words •Negative emotions •Anger •Death	•Articles
Conscientiousness	•Achievement	•Exclusive •Negations •Negative emotions •Anger •Body •Death •Swear words	
Emotional stability	•Word count •Positive emotions	•Inclusive •Negative emotions •Anxiety •Conjunctions	•Humans •Work
Openness to experience	•Perceptual processes •Death •Articles •Prepositions	•Social processes •Family •First person singular pronoun •Past tense verbs	•Positive emotions •Grooming

Note: Variables and relationship direction were listed only if at least two studies found that LIWC variable to be statistically significantly related to that trait. Conflicting findings lists LIWC variables found to be significantly positively *and* negatively related to a trait. We used Qiu et al. (2012), the studies listed in their Table 1, and the Kern et al. (2014). For large-scale studies (e.g., Kern et al., 2014; Yarkoni, 2010), we only included categories significant at $p < .01$. Otherwise, we used $p < .05$.

Table 2

Means, standard deviations, and correlations between gender, interviewer-rated traits, and self-reported traits

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Gender	.50	.50	--													
Interviewer-report																
2. Extraversion	4.53	1.14	-.14**	(.89)												
3. Agreeableness	4.88	.67	-.33**	.31**	(.69)											
4. Conscientiousness	5.55	.58	-.01	.28**	.16**	(.73)										
5. Emotional Stability	5.19	.58	.05	.37**	.27**	.39**	(.66)									
6. Openness	4.58	.87	.02	.42**	.14**	.38**	.32**	(.79)								
Self-report																
7. Extraversion	3.18	.85	-.13*	.41**	.15**	.03	.19**	.07	(.90)							
8. Agreeableness	4.02	.59	-.31**	.29**	.46**	.05	.14**	.14**	.30**	(.83)						
9. Conscientiousness	3.55	.63	-.12*	.07	.05	.06	.10*	-.16**	.11*	.14**	(.80)					
10. Emotional Stability	3.01	.73	.17**	-.04	-.03	-.01	.12*	-.05	.14**	.03	.14**	(.84)				
11. Openness	3.68	.55	.11*	.08	-.00	.02	.04	.16**	.21**	.25**	.11*	.04	(.76)			
12. HS GPA	3.75	.25	-.19**	.14*	.20**	.14*	.05	.10	-.04	.14*	.23**	.02	.11*	--		
13. SAT Verbal	619	102	.11*	.03	-.03	.10	-.05	.12*	-.10	-.10	-.01	-.07	.17**	.16*	--	
14. SAT Math	637	121	.29**	-.08	-.20**	.14*	-.07	.12*	-.19**	-.26**	-.03	-.02	.04	.15*	.67**	--
15. ACT	27.4	4.31	.03	.14*	-.02	.18*	.13	.29**	-.07	-.02	.03	.03	.11	.26**	.31**	.29**

Note. M = mean. *SD* = standard deviation. * indicates $p < .05$. ** indicates $p < .01$. HS GPA = high school grade point average. Reliabilities reported in diagonal. Interviewer reliabilities are ICC(C, 8), and interviewee self-reported reliabilities are Cronbach's alpha. Self-reports (N = 389) were scored on a five-point scale, and interviewer ratings (N = 441) were made on a seven-point scale. For gender, female=0 and male=1. HS GPA N = 383. SAT N = 313. ACT N = 230.

Table 3

Significant correlations between reported personality traits and LIWC variables

Self-reports						Interviewer-reports				
E	A	C	ES	O	LIWC variable	E	A	C	ES	O
.11	.14				1st per. sing. pronouns					
					2nd per. pronouns					-.14
	.11				3rd per. plur. pronouns		.13			
	.14				Affective processes		.23	-.18		-.10
					Anger		-.13			
					Anxiety	-.12				
	-.11				Articles		-.16			.11
					Assent	-.19		-.29	-.20	-.14
		-.13			Auxiliary verbs	.12				
	.10				Certainty		.15	-.11		-.10
	.11				Cognitive processes	.11	.12			
					Common adjectives		.12			
	.10				Common adverbs	.13	.13			
		-.16			Common verbs	.13	.10			
	.17				Conjunctions	.19	.18	.10		
.10	.14				Differentiation	.16	.13			
	.11				Discrepancy		.11			
					Drives		.17			-.14
					Drives: affiliation		.14	-.10		
		.11	.12		Drives: power					-.19
		.10			Drives: rewards		.22			-.15
				-.15	Family		.11	-.17		-.14
				-.13	Feel		.15			
					Female references		.14			
.13	.20				Function words	.27	.20			
					Hear				-.10	.13
					Health		.12			
					Home		.17	-.12		
					Impersonal pronouns	.14	.11			
			.11		Informal	-.18		-.26	-.18	-.10
					Ingestion	-.13	-.10	-.09	-.14	-.10
					Insight	.10		.12		.10
		-.11			Interrogatives					
.12		.12		-.16	Leisure			-.14		-.14
	-.11				Money					
.11					Motion words				.10	

Table 3 (*continued*)

Self-reports						Interviewer-reports				
E	A	C	ES	O	LIWC variable	E	A	C	ES	O
					Negative emotions	-.16		-.11	-.10	-.14
	-.14				Numbers	-.14	-.23		-.11	-.10
			-.11		Past focus	.10				.10
					Perceptual processes					.10
.14	.19				Personal pronouns		.14		-.10	
.12	.19				Positive emotions		.27	-.15		
-.12					Prepositions			.14		
					Present focus			-.10		-.12
.15	.21				Pronouns	.16	.18			
					Quantifiers	-.12		-.11	-.14	
	.12				Religion		.17			
					Sad			-.12	-.13	-.12
					See			-.12		
			-.12		Sexual					
.12	.15				Social processes		.23	-.14		-.10
-.10					Tentative					
					Time	-.11				
.13					Word count	.45	.13	.31	.13	.29
		.10			Words > 6 letters		-.12	.23	.17	.16
		.13			Work		-.13	.15		

Note: All correlations significant at $p < .05$. Non-significant correlations suppressed for readability.

Table 4

10-fold cross-validated accuracy for predicting interviewee personality traits (automatic personality recognition & perception)

	Elastic Net Parameters		Convergent Correlations				
Self-reports	Alpha	Lambda	$\bar{r}$	r_{min}	r_{max}	r_{SD}	$\bar{\rho}$
Extraversion	1.0	.0000619	.27	.10	.45	.11	.29
Agreeableness	.1	.1044	.25	-.10	.48	.16	.27
Conscientiousness	.9	.0951	.17	-.12	.47	.18	.19
Emotional Stability	.3	.0362	.13	-.22	.29	.19	.14
Openness to Experience	.9	.0121	.12	-.07	.46	.15	.14
		AVERAGE:	.19	-.08	.43	.16	.20

Interviewer-reports	Alpha	Lambda	$\bar{r}$	r_{min}	r_{max}	r_{SD}	$\bar{\rho}$
Extraversion	.2	.449	.49	.37	.61	.08	.52
Agreeableness	.1	.0668	.46	.23	.66	.14	.55
Conscientiousness	.9	.0675	.41	.09	.64	.18	.48
Emotional Stability	.1	.102	.21	-.00	.38	.15	.26
Openness to Experience	1.0	.0404	.39	.15	.56	.12	.44
		AVERAGE:	.39	.17	.57	.13	.45

Note: Hyperparameters reported for the most accurate models. $\bar{r}$ calculated by correlating predicted and reported traits in each fold, converting r to Fisher's z, averaging z across the 10 folds, then converting $\bar{z}$ to $\bar{r}$. r_{min}, r_{max} = minimum and maximum convergent correlations, respectively. r_{SD} = standard deviation of the convergent correlations. ρ = average correlation corrected for self- or interviewer-rating unreliability.

Appendix A: Hierarchical Regression Using Self-Reported and Interviewer-Reported Personality to Predict Academic Outcomes

Table A1

Regression analysis of high school GPA, SAT verbal scores, SAT math scores, and ACT scores beginning with self-reports

	High School GPA (N = 383)		SAT Verbal (N = 313)		SAT Math (N = 316)		ACT (N = 230)	
Variable/Step	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Intercept/Constant	-.00 (.05)	-.01 (.05)	-.01 (.06)	-.01 (.05)	-.01 (.05)	-.02 (.05)	-.01 (.07)	-.03 (.07)
Gender	-.19 (.05)**	-.14 (.0)*	.05 (.06)	.08 (.06)	.21 (.06)**	.22 (.06)**	-.01 (.07)	-.02 (.07)
Self-reported Extraversion	-.13 (.05)*	-.16 (.06)**	-.10 (.06)†	-.11 (.06)†	-.12 (.06)*	-.11 (.06)†	-.11 (.07)	-.15 (.08)*
Self-reported Agreeableness	.07 (.06)	-.01 (.06)	-.10 (.06) †	-.13 (.07)†	-.17 (.06)*	-.15 (.06)	-.01 (.07)	-.02 (.08)
Self-reported Conscientiousness	.20 (.05)**	.22 (.05)**	.02 (.06)	.04 (.06)	.03 (.05)	.06 (.06)	.04 (.07)	.09 (.07)
Self-reported Emotional Stability	.04 (.05)	.06 (.05)	-.08 (.06)	-.05 (.06)	-.05 (.06)	-.03 (.06)	.03 (.07)	.04 (.07)
Self-reported Openness	.12 (.05)*	.12 (.05)*	.21 (.06)**	.20 (.06)**	.09 (.06)	.05 (.06)	.13 (.07)†	.11 (.07)
Interviewer-rated Extraversion		.10 (.06)		.06 (.07)		-.00 (.07)		.11 (.08)
Interviewer-rated Agreeableness		.15 (.06)*		.05 (.07)		-.06 (.06)		-.07 (.08)
Interviewer-rated Conscientiousness		.09 (.06)		.10 (.07)		.19 (.06)**		.07 (.08)
Interviewer-rated Emotional Stability		-.09 (.06)		-.11 (.07)†		-.14 (.06)*		.02 (.08)
Interviewer-rated Openness		.06 (.06)		.08 (.07)		.15 (.06)*		.25 (.08)**
Total R^2	.11	.16	.07	.10	.13	.19	.02	.13
ΔR^2		.05**		.03		.06**		.11**

Note: † $p < .1$. * $p < .05$. ** $p < .01$. Standard errors in parenthesis.

Table A2

Regression analysis of high school GPA, SAT verbal scores, SAT math scores, and ACT scores beginning with interviewer-reports

	High School GPA (N = 383)		SAT Verbal (N = 313)		SAT Math (N = 316)		ACT (N = 230)	
Variable/Step	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Intercept/Constant	-.01 (.05)	-.01 (.05)	-.00 (.06)	-.01 (.05)	-.01 (.05)	-.02 (.05)	-.01 (.06)	-.03 (.07)
Gender	-.13 (.05)*	-.14 (.05)**	.13 (.06)*	.08 (.06)	.26 (.06)**	.22 (.06)**	.01 (.07)	-.02 (.07)
Interviewer-rated Extraversion	.05 (.06)	.10 (.06)	.01 (.07)	.06 (.07)	-.06 (.06)	-.00 (.07)	.05 (.08)	.11 (.08)
Interviewer-rated Agreeableness	.14 (.06)*	.15 (.06)*	.01 (.06)	.05 (.07)	-.11 (.06)†	-.06 (.06)	-.09 (.07)	-.07 (.08)
Interviewer-rated Conscientiousness	.12 (.06)*	.09 (.06)	.12 (.07)†	.10 (.07)	.22 (.06)**	.19 (.06)**	.09 (.08)	.07 (.08)
Interviewer-rated Emotional Stability	-.07 (.06)	-.09 (.06)†	-.14 (.07)*	-.11 (.07)†	-.14 (.06)*	-.14 (.06)*	.02 (.08)	.02 (.08)
Interviewer-rated Openness	.04 (.06)	.06 (.06)	.11 (.06)†	.08 (.07)	.15 (.06)*	.15 (.06)*	.26 (.07)**	.25 (.08)**
Self-reported Extraversion		-.16 (.06)**		-.11 (.06)†		-.11 (.06)†		-.15 (.08)*
Self-reported Agreeableness		-.01 (.06)		-.13 (.07)†		-.15 (.06)*		-.02 (.08)
Self-reported Conscientiousness		.22 (.05)**		.04 (.06)		.06 (.06)		.09 (.07)
Self-reported Emotional Stability		.06 (.05)		-.06 (.06)		-.03 (.06)		.04 (.07)
Self-reported Openness		.12 (.05)*		.20 (.06)**		.05 (.06)		.11 (.07)
Total R^2	.08	.16	.05	.10	.16	.19	.10	.13
ΔR^2		.08**		.05**		.03*		.03

Note: † $p < .1$. * $p < .05$. ** $p < .01$. Standard errors in parenthesis.