Can Psychological Traits Be Inferred from Spending?

Evidence from Transaction Data

Joe. J. Gladstone*, Sandra C. Matz* & Alain Lemaire

*Both authors contributed equally; author order is alphabetical.

Author Notes:

Joe J. Gladstone (Corresponding Author)

UCL School of Management

1 Canada Square

London E14 5AA

j.gladstone@ucl.ac.uk

+ 44 (0) 20 3108 6095

Sandra C. Matz

Columbia Business School

3022 Broadway

New York, 10027

sm4409@gsb.columbia.edu

+1 (646) 647-5341

Alain Lemaire

Columbia Business School

3022 Broadway

New York, 10027

alemaire18@gsb.columbia.edu

ABSTRACT

The automatic assessment of psychological traits from digital footprints allows researchers to study psychological traits at scale and in settings of high ecological validity. In this research, we investigate whether spending records – a ubiquitous and universal form of digital footprint – can be used to infer psychological traits. We apply an ensemble machine-learning technique (Random Forest) to a dataset combining two million spending records from bank accounts with survey responses from the account holders (N = 2,193). Our predictive accuracies are modest for the Big Five personality traits (r = 0.15, corrected ρ = 0.21), but provide higher precision for specific traits, including Materialism (r = 0.33, corrected ρ = 0.42). We compare the predictive accuracy of these models with alternative digital behaviors used in past research, including those observed on social media platforms, and show that the predictive accuracies are relatively stable across socio-economic groups and over time.

Keywords: Big Five personality, consumer psychology, psychometrics, computational social science, financial decision-making.

The automatic prediction of psychological traits from digital footprints offers the potential to transform the scientific investigation of individual differences, by allowing researchers to study psychological traits at scale and in settings of high ecological validity. Driven by advances in computational methods and the wider availability of user-generated data (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015), this research suggests that psychological traits can be inferred from digital records of behavior, including from Facebook profiles (Park et al., 2014; Youyou, Kosinski, & Stillwell, 2014), Twitter (Golbeck, Robles, Edmondson, & Turner, 2011), Flickr pictures (Segalin, Perina, Cristani, & Vinciarelli, 2017) and even music collections (Nave et al., 2018).

We extend this research by investigating whether patterns in a person's *spending* can also reveal differences in psychological traits. After all, spending is often considered a reflection of who we are as individuals (self-congruity theory, Sirgy, 1985): We buy products not only for what they can do, but also for what they mean to us (Levy, 1959). With more than 14 billion payment cards in circulation[1], aggregated spending records provide a detailed metric of people's tastes and preferences, and this, combined with evidence that categories of spending have consistent associations with personality (Matz, Gladstone, & Stillwell, 2016), suggests that it may be possible to accurately infer a person's psychological profile using their spending records.

However, it is unclear how the accuracy of spending records will compare to that obtained with other types of digital footprints. In fact, there are competing hypotheses that can be derived from the theoretical distinction of two types of observable behavioral traces, both of which have been shown to hold valid cues to a person's psychological profile: identity claims and behavioral residues (Gosling, Ko, Mannarelli, & Morris, 2002). Social media platforms are designed as a way

[1] World Payments Report 2017 (WPR 2017). Retrieved from: www.worldpaymentsreport.com

for individuals to explicitly communicate their identity and preferences to others within their social network. As such digital traces observed on social media can be thought of as “identity claims”: deliberate symbolic statements made by individuals to express themselves (Gosling et al., 2002). Digital music libraries or spending records, on the other hand, may be better described as “behavioral residues”: subtle cues about people’s preferences inadvertently conveyed as a result of one’s activity (Gosling et al., 2002). On the one hand, because spending is recorded passively and includes information often hidden from others, it might be less influenced by social desirability and therefore a more accurate reflection of a person’s psychological traits. On the other hand, spending records, like other behavioral residues, may be less predictive of people’s psychological traits because they are weaker signals of how a person both perceives themselves and desires to be perceived by others.

We explore this question by comparing the predictive accuracy of models built using spending records with alternative digital behaviors used in past research. We also extend past research by investigating whether predictive accuracies from models built from spending records are biased against certain groups, such as those who are poorer (and therefore spend less).

METHOD

To investigate whether personality can be predicted from a person’s spending, we combined two million spending records from bank accounts with survey responses completed by the account holders (*n* = 2193). Using a machine-learning “predictive” approach (Yarkoni & Westfall, 2017), we examined the accuracy of out-of-sample predictions of participants’ personalities, and provide comparative estimates for how well different traits can be predicted from spending records. All customer data was fully anonymized, and we received ethical approval for the project from the university ethics committee.

Participants

The dataset was collected in collaboration with a UK-based money management app in May 2017. The service provides users with a single dashboard of their finances by aggregating transactions from across all their bank accounts and providers. For example, if a participant had two credit cards and one checking account, each with different service providers, then data from each of these accounts will be recorded by the application. This pooling of account information represents an advantage over previous research using bank account data which has typically relied on information derived from only a single bank (e.g., Matz et al., 2016).

Customers of the service were sent a survey link by email asking them to take part in the study, with the opportunity to win a tablet computer as a prize. Within the survey, participants consented to match their survey responses with their transaction data for research purposes. In total, 2,193 people completed the study and provided their consent to participate. For 1,875 of those participants, the service provided information on age (M = 38.07 years, SD = 11.46) and annual income. Gender was not measured directly but derived by running first names of account users through a names database, providing gender in just over half of cases (11% female, 43% male, 46% unknown).

Survey Measures

While past research has focused primarily on using the Big Five personality traits (Costa & McCrae, 1992), in this project we chose to additionally predict two traits that research has found to be more closely related to consumption: materialism and self-control. For example, materialistic people prefer material goods over experiential ones (Howell, Pchelin, & Iyer, 2012), and those with greater self-control spend less on impulsive purchases and save more (Oaten & Cheng, 2007).

Thus, we expected spending records to provide more accurate predictions for these more specific and relevant traits, compared to the broader Big Five dimensions.

Big Five personality. The most widely accepted model of personality, the ‘Big Five’ model (Costa & McCrae, 1992), proposes a taxonomy of five personality traits: Openness (open to new experience, complex vs. conventional, uncreative), Conscientiousness (dependable, self-disciplined vs. disorganized, careless), Extroversion (extroverted, enthusiastic vs. reserved, quiet), Agreeableness (sympathetic, warm vs. critical quarrelsome) and Neuroticism (anxious, easily upset vs. calm, emotionally stable). To measure each trait, we used the BFI-10 instrument, an established short scale of the Big Five model (Rammstedt & John, 2007). The correlations between trait item-pairs ranged between $r = 0.20$ for Agreeableness and 0.59 for Extroversion. With Cronbach’s alphas ranging from $\alpha = 0.33$ for Agreeableness to $\alpha = 0.75$ for Extroversion, the internal consistencies of scales were found to range from poor to acceptable. Given the relatively low internal consistencies of our outcome measures, we accounted for measurement error in our analyses by correcting for attenuation.

Materialism. Materialism refers to the degree to which an individual considers material possessions and physical comfort important, and was measured using three items taken from a widely used measure of materialism (Richins & Dawson, 1992): (1) “I admire people who own expensive homes, cars and clothes”, (2) “I like a lot of luxury in my life”, and (3) “I’d be happier if I could afford to buy more things”. With a Cronbach’s alpha of $\alpha = 0.62$ the scale reliability was found to be low. Similar to the Big Five, we also correct for attenuation in our analysis.

Self-control. Self-control refers to the degree to which individuals can control and regulate their impulses. The construct was measured using a single item (“I am good at resisting temptation”) from the Brief Self-Control Scale (BSCS; Tangney, Baumeister, & Boone, 2004).

Given that the nature of our data collection required us to limit the number of questions to a bare minimum, we believe that – although not ideal - the single item measure provides a pragmatic approximation of the wider construct of self-control.

Spending Records

The transaction records obtained from participants' banking data encompassed detailed information of all purchases made using customers' bank accounts (checking accounts and credit cards) over a period of twelve months.

Broad spending categories. Individuals' purchases were automatically grouped by the company into 279 categories, including 'supermarkets', 'furniture stores' and 'insurance policies'. The full list of spending categories with the average annual amount spent and number of purchases per category is provided in Table S1 in the online supplementary material. For the purpose of our analysis, we aggregated participant's spending in each of the categories over the year preceding the survey.

We used participant's relative spending across categories, rather than the raw amounts, to help ensure our predictors reflected patterns in spending, rather than simply the participant's total income or wealth. To calculate this, we divided participant's spending in an individual category by their overall spending, giving us the relative amount spent in each specific category[2]. We then centered each of the 279 features before using them as features in the prediction models. It is worth noting that because we expect wealthier individuals to spend a lower proportion of their

[2] For example, if a participant had spent a total of £500 on books, and £30,000 overall, the relative spending on books would be 1.7%, the same as for a person who spent only £50 on books and £3,000 overall.

overall budget on necessities, our proportional measure of spending might still vary across the resource levels of the participants (we test the implications of this on the accuracy of our models more explicitly in the context of moderation analyses).

Specific merchant names. Each transaction was also associated with a specific merchant name (e.g., Tesco, or Amazon). Overall, customer purchases were tagged with one of 658 identifiable merchants, or were given a "no merchant tag" reserved for all transactions with merchants that could not be identified. The merchants that were represented most frequently in our dataset were Tesco (UK supermarket, 5.68%), Amazon (online retailer, 4.01%), Sainsbury (UK supermarket, 3.80%) and PayPal (money transfer service, 3.36%).

To reduce the dimensionality of the merchant tag data, we applied Latent Dirichlet Allocation (LDA), an established technique within probabilistic machine learning used to cluster data. While LDA is most commonly used in classifying text, it can also be used in contexts such as ours, where the goal is to group data based on shared characteristics. LDA assumes that there is a combination of merchant tags that frequently appear together and form a coherent theme or topic. In our data, for example, the merchants McDonald's, KFC, Pizza Hut, Subway and Burger King formed together into a single topic (which we label 'Fast Food'), whereas Costa, Starbucks and Caffe Nero formed into a separate topic (which we label 'Coffee Shops'). LDA then assumes that the spending habits of each person can be described as a weighted combination of these themes. For example, one person's transactions could be a combination of 80% fast food and 20% coffee, while another person's spending could be made up of 50% each.

To apply LDA to our data, we aggregated all the identifiable merchant tags used by each customer during our observation period into a "bag of tags"[3]. Following the recommended procedure for identifying the optimal number of topics, we split the bags of tags into a standard 80-20 training and test set and used the perplexity score – an out-of-sample measure of model fit – as the criterion to evaluate each model. The analyses yielded a total of 34 topics in our dataset. As well as the Fast Food and Coffee Shop topics mentioned previously, other topics we identified related to Investment Services, Utilities and Electronics (see OSF link for an interactive chart of topics https://tinyurl.com/y8fw4b4p). These 34 topics, as well as the 279 broad spending categories, were included as features in our predictive model.

Prediction Models

Machine learning algorithms provide new opportunities for researchers in psychology to gain valuable insights from large-scale behavioral data. While the traditional psychological toolbox has provided tools well-calibrated to analyze the results from classic experimental paradigms (e.g. when comparing a limited number of experimental conditions), it has yet to expand to provide appropriate tools for analyzing large-scale observational datasets of human behavior ("Big Data"). In order to predict psychological traits from participants' transaction histories, we employed random forest models. These models are widely-used in other disciplines, including computer science research, but remain rare in psychological research. Thus, in order to make the method accessible to our readers – and to encourage other researchers to use these methods in their own

[3] If a participant purchased twice at Amazon.com and once at Apple, the individual's bag of tags would be "Amazon, Apple, Amazon".

work – we first provide a brief explanation of random forest models, and subsequently describe how we apply the model in the context of our data.

An Introduction to Random Forest Models

To grasp the logic behind random forest models, it is necessary to start with an explanation of their essential building blocks: decision trees. Although we rarely label them as such, decision trees are simply a way of describing the logic we use to make decisions on a daily basis. Take a clinical psychologist, for example, who is trying to diagnose a patient suffering from a list of symptoms. To recommend appropriate treatments, the psychologist needs to determine whether their patient is depressed, and if so, how severely. To classify the patient as mildly, severely or not depressed at all, the psychologist asks questions such as "Have you lost interest in activities which you used to enjoy?". These questions form the branches of the tree, and as the patient answers each of these questions, the therapist gets closer to her diagnosis. If the patient confirms they have lost interest in enjoyable activities, then the likelihood that the patient is suffering from depression, for example, increases. Based on this answer, the therapist might decide to follow up with a further question, such as "Have you had suicidal thoughts?", aimed at confirming or disconfirming the initial hypothesis. After going through several of these decision trees, the therapist makes their diagnosis: the patient is indeed depressed, and severely so.

Just like their real-world counterparts, decision trees in random forest models are aimed at narrowing down the set of possible outcomes. When trying to predict a person's extroversion from their transaction data, for example, the model could ask whether that person spends a large amount of money on dining and drinking, a spending category found to be correlated with Extroversion (Matz et al. 2016). If the answer is *yes* then the likelihood of the participant being more extroverted than the average person increases – if the answer is *no* the likelihood decreases.

While the psychologist in our example decides upon their questions a priori, by using the DSM-5 manual, the "questions" our random forest model asks are instead developed by the model based on the data used to train it. In our case, the model learns to map input data (spending records) to output data (personality scores) by learning how to build a decision tree that yields the highest accuracy in mapping one to the other. This can be compared to learning the optimal questions to ask to diagnose depression.

Random *forest* models are a combination of large numbers of decision *trees*. To explain why we need to combine trees together into a forest, let us return to our therapist who has diagnosed their patient as severely depressed. Was this diagnosis the correct one? Even with highly standardized diagnostic tools, there remains ample room for subjectivity and error in these types of decisions. Indeed, over 60% of patients diagnosed with depression by a clinician did not meet the official criteria for the disorder upon re-evaluation (Mojtabai, 2013). Therefore, should we trust the opinion of this one therapist, or instead get a second opinion? Better yet, we could even source diagnoses from a few hundred or thousand therapists. In the latter case, some therapists will diagnose the patient as mentally healthy, while others will diagnose her as severely depressed. By relying on the diagnosis most commonly suggested across this pool of psychologists, we are more likely to end-up with the correct diagnosis, compared with relying on any individual therapist chosen at random. This simple concept of "the wisdom of the crowds", where groups of people pool their abilities to show collective intelligence, is what underlies the predictive accuracy of random forest models, where a large number of decision trees are aggregated together to improve predictive accuracy.

Finally, what makes random forests *random*? Each decision tree in a random forest model has access only to a random subset of input data (e.g. spending categories) and a random subset of participants. This artificial increase in variance allows the final model to produce a more robust mapping of input and output data. This is similar to our clinical psychologist being provided with only a random subset of questions to ask their patients, and being able to draw only on a random subset of prior experiences to make her diagnosis.

Model Specifications

Based on the logic of random forest models outlined above, the model we fitted to our data has the following two features. First, each decision tree is constructed using a different bootstrapped subsample of the data. Second, the split of each node in a tree is determined by a random subset of predictors. Combined, these two features make random forest models particularly robust to overfitting.

To further reduce the risk of overfitting, we followed a standard 10-fold cross-validation protocol. Cross-validation involves training and testing a model on different samples of data, allowing us to quantify the out-of-sample prediction error. This means we first randomly split the dataset into ten samples, and then train the random forest model using 90% of the dataset (nine training samples) to predict the Big Five traits, materialism and self-control. This training sample is then separated into a 66% fitting sample and a 33% validation sample. Using the fitting and validation set, we performed an exhaustive grid search over the parameters of the Random Forest to determine the optimal model specification. During the grid search process, we varied the following three parameters: 1) the number of trees was varied from 50 to 350 in increments of 50; 2) the maximum depth of the trees was varied between 5, 15, 30, 60 and 120; and 3) the number of predictors to consider at each split were varied from the log(k), $\sqrt{k}$, and k, where k is the number of predictors. In a second step, we used the best trained model to predict the scores of participants in the remaining 10% of the dataset (the hold-out testing sample). In a third step, we estimate the predictive accuracy of our model by calculating Pearson product-moment correlations between the predicted and actual scores for the Big Five traits, materialism and self-control. This three-step procedure is repeated 10 times, each time with different data used in the training and testing data sets.

RESULTS

The three central questions motivating our research are: (i) Can spending records predict a person's psychological traits? (ii) Do socio-demographic variables moderate the predictive accuracy of spending records? and (iii) How does the predictive accuracy of spending records compare to other digital footprints? We present our results in response to these questions.

Can spending records predict psychological traits?

Across all psychological traits measured in our study, the average correlation between actual and predicted scores was $r = 0.19$. However, this aggregated measure of accuracy masks considerable variation across individual traits, with Openness having the lowest accuracy ($r = 0.12$), and Materialism having the highest ($r = 0.33$). The differences in predictive accuracy across the different traits suggests that transaction records provide greater predictive accuracy for more focused psychological traits (materialism and self-control) than participant's more general psychological traits (Big Five personality traits). While the average accuracy for the narrow traits was $r = 0.30$, the average accuracy for the broad personality traits was only $r = 0.15$. If we correct for attenuation, a procedure that accounts for measurement error in the outcome variable, the average correlation for all traits increases to $\rho = 0.24$ (see Table S2 in the supplementary material lists the individual correlations, including after they have been corrected for attenuation).

To develop a more intuitive understanding of which spending categories were driving the predictive accuracy in each of the models, we calculated the univariate correlations between each of the spending categories and each of the psychological traits. Table S3 in the online supplementary material lists the five categories most positively and negatively correlated with each trait. For example, openness was found to be positively related to spending money on “flights”, extroversion with “dining and drinking”, agreeableness with “donations”, conscientiousness with “savings”, and materialism with “jewelry”. Similarly, there was a negative association found between self-control and “bank charges”, materialism and “donations” and neuroticism with “mortgage payments”. The direction of these correlations provide some face validity for the expected relationships between categories and personality, supporting prior associations found in consumer psychology, such as conscientious individuals allocating more money to savings and investments, open-minded individuals spending more money on travel, and materialistic individuals giving less to charity (Belk, 1985; Matz et al., 2016; Mosca & McCrory, 2016).

Returning to the overall predictive accuracies, we further explored the degree to which the predictive accuracy of our models depends on the amount of data available about a participant. In other words, would our predictions remain relatively stable if we had collected only a single month of data about a person, or do the random forest models require the full year of transaction data? To answer this, we re-calculated our models by training data on 90% of participants and predicting the scores of the remaining 10% using 1 to 11 months of their data. This process was repeated 10 times, such that the resulting accuracies represent 10-fold cross-validated averages. Overall, the analysis suggests that the accuracy of the models increases after being supplied additional months of data. For example, the average predictive accuracy across all traits after one month ($r = 0.15$) was lower than after six months ($r = 0.19$). However, the stability of these associations also shows that a degree of accuracy can be obtained with a relatively limited amount of data about individuals (see Figure S1 in the supplementary online material).

For the purposes of comparison, we also built a model using the demographic information we had available for each user as predictors; their age and gender. This provided an average correlation of $r = 0.11$ across all traits ($\rho = 0.17$), illustrating that spending accounts for a greater proportion of the variance in personality compared to a person's age or gender combined. However, it should be noted that while demographics were less predictive of personality than spending records across all traits, this was not universally the case; they were more accurate than spending records when predicting the Big Five trait of Agreeableness. When combined in a single model, demographics and spending records predicted the psychological traits with an average accuracy of $r = 0.20$ ($\rho = 0.27$), outperforming the predictions of either spending records or demographics alone. One can think of this approach as a way to "norm" our prediction models, just like self-report questionnaires often provide separate norms for different socio-demographic groups (e.g. age and gender). For example, spending more money on jewelry might be predictive of higher levels of materialism in women, while it might instead be predictive of higher agreeableness in men (who are more likely to buy jewelry for their loved ones). Adding age and gender as features allows the random forest model to use those characteristics in building the trees and hence to develop idiosyncratic models for each socio-demographic group. The predictive validity of these models is provided in Table S3 and Figure S2 in the online supplementary materials.

Do socio-demographic variables moderate the predictive accuracy of spending records?

An important question when it comes to the accuracy of predictive algorithms is whether they can predict outcomes with a similar degree of accuracy across different socio-demographic groups. For example, it is possible that an algorithm which predicts personality from spending is more accurate for those who spend more overall, because these people have a greater number of purchases to evaluate, and they may also have more discretion over the purchases they make. To explore this question, we test whether the absolute error of our predictions – that is the absolute difference between predicted and actual scores – depends on the following socio-demographic variables: age, salary (logged), total spending (logged), and deprivation level (logged). While age, salary and total spending were provided directly by the app, we calculated the deprivation level of their local area by matching participants' postal areas with UK census data. As such the deprivation variable approximates a participant's deprivation level by using the deprivation level of their residential neighborhood.

To calculate a single measure of prediction error, we average the prediction error across the seven predictions we make for each person (five personality traits, materialism and self-control). The results of a linear regression analyses suggest that, overall, the predictive accuracy is relatively stable across the socio-demographic variables we investigate. With an adjusted R^2 of 0.004, our moderators explain less than 1 percent of the variance in prediction error. The only variable that became significant at an alpha level of 0.05 was deprivation ($B = 0.25$, $SE = 0.11$, $\beta = 0.058$, $p = 0.028$), with participants living in areas that are highly deprived being more difficult to predict (see Table S4 in the supplementary material for the full model output). In order to develop a better understanding of how deprivation relates to predictive accuracy, we ran additional regression analyses predicting the absolute difference for each trait from the deprivation index. While the effect goes in the same direction for all traits, it reached significance only for Neuroticism ($B = 0.55$, $SE = 0.25$, $\beta = 0.072$, $p = 0.029$) and Self-control ($B = 0.33$, $SE = 0.14$, $\beta = 0.061$, $p = 0.018$). We visualize these effects in Figure 1, using surface level plots, which illustrate the relationship between actual and predicted scores on these two traits as a function of deprivation.

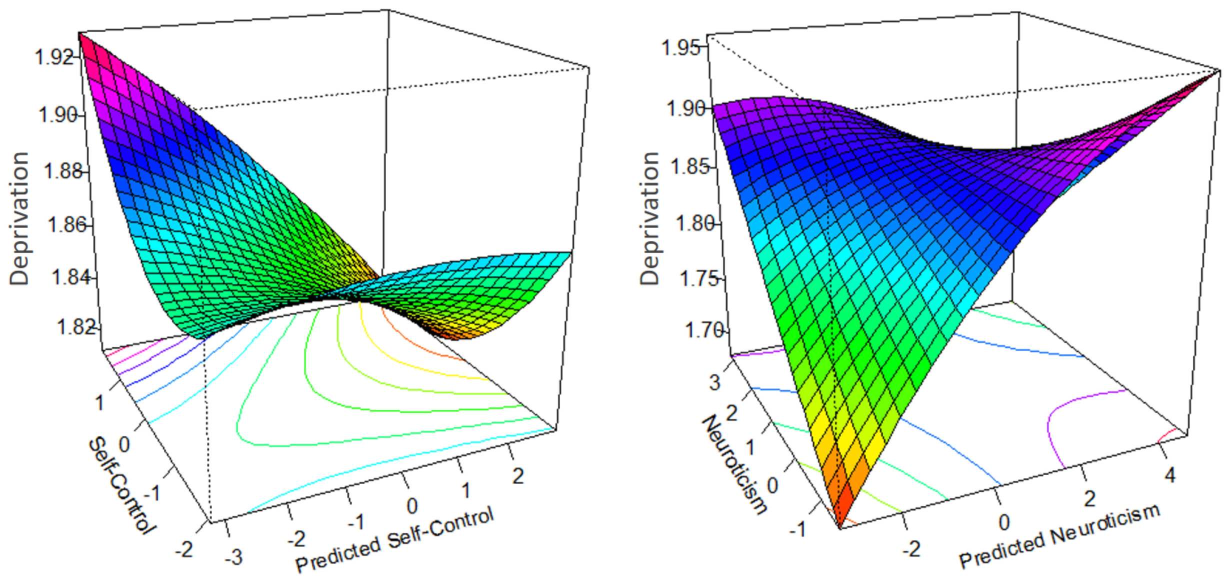
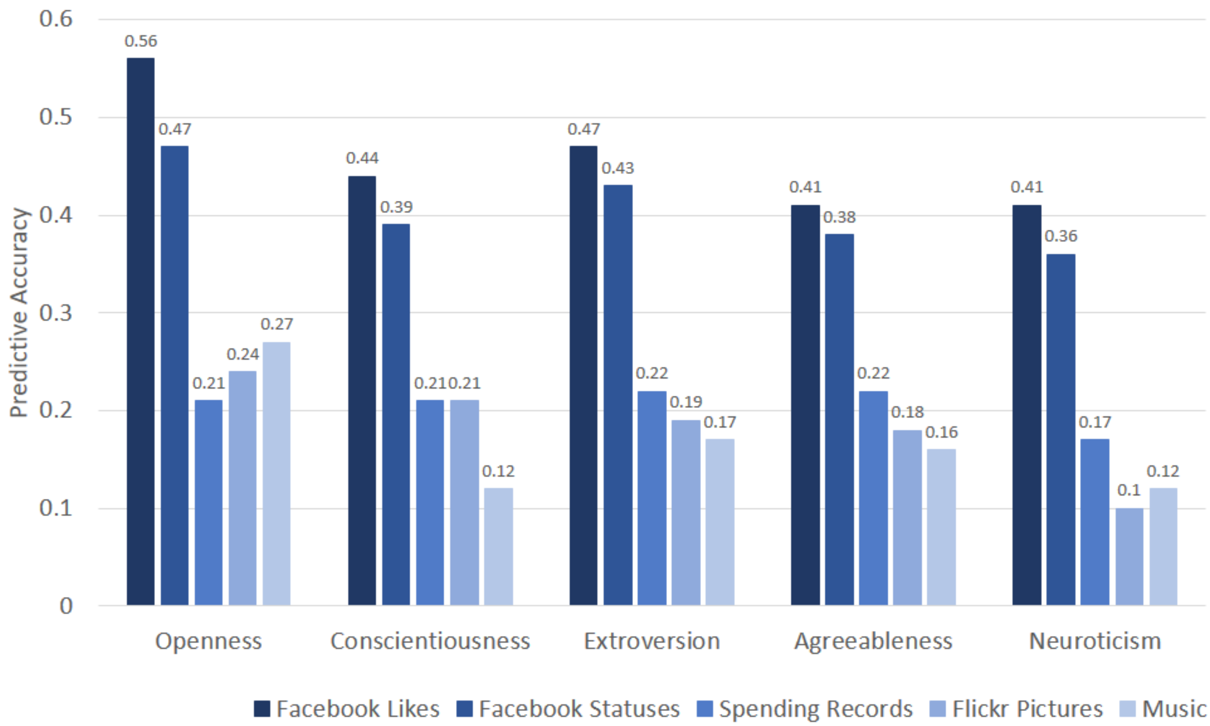


Figure 1. Surface-level plots illustrating the moderating effect of postcode level deprivation on the predictive accuracy of our model for Self-Control (left) and Neuroticism (right).

One possible explanation for the moderating effect of postcode-level deprivation is that people living in poorer neighborhoods are less likely to have a large discretionary spending budget to allocate to products and services which align with their personality. Instead they may need to use their resources for essential necessities (e.g. groceries). However, given we did not find significant effects for either total spend or income at the individual level, the findings might instead suggest that deprived areas offer less opportunities to spend money in a way that reflects psychological preferences. However, it should be noted that these differences are both practically small and statistically weak, and thus future research should aim to replicate these findings to ensure they are robust before greater emphasis is given to interpreting their meaning.

How does the predictive accuracy of spending records compare to other digital footprints?

To illustrate the relative accuracy of these prediction models, and to contextualize them within the broader literature of automatic personality prediction, in Figure 2 we compare our results with those reported in recent studies predicting the Big Five traits from several online digital footprints, including Facebook Likes (Youyou et al., 2014), Facebook status updates (Park et al., 2014), Flickr pictures (Segalin et al., 2017) and music preferences (Nave et al., 2018). While the predictions from Facebook Likes and status updates outperform our predictions considerably, our findings are comparable to those found in the context of music preferences and online photo-sharing websites.



Personality measure (# items)	IPIP (20-100)	IPIP (20-100)	BFI-10 (10)	BFI-10 (10)	IPIP (20-100)
Sample size	86,220	71,556	2,193	300	22,252
Model	Lasso	Ridge	Lasso	Lasso	Lasso
Model tuning	Yes	Yes	Yes	Yes	Yes
Dimensionality reduction (Y/N)	No	Yes	Partly	Yes	No

Figure 2. Predictive accuracies of the Big Five traits comparing Facebook likes, Facebook statuses, spending records, Instagram pictures and music preferences. Bars represent correlations accounting for attenuation which we calculated based on the alphas reported in the original manuscripts or provided by the authors. Facebook likes (Youyou et al., 2014), Facebook status updates (Park et al., 2014), Spending records from this study, Flickr pictures (Segalin et al., 2017) and music preferences (Nave et al., 2018).

These results are consistent with the distinction between “identity claims” and “behavioral residues” we outlined in the introduction (Gosling et al., 2002). Facebook is designed for users to communicate their preferences and express themselves to others and therefore predominantly captures identity claims. In contrast, digital music libraries or spending records are far less curated and hence constitute more subtle behavioral residues. Although these comparisons of accuracies across studies are far from conclusive (as Figure 2 illustrates, the studies differ on a number of dimensions), it provides suggestive evidence that behavioral residues might be less predictive of people's self-reported personality than identity claims. In addition to providing a weaker social signal, the lower accuracy of spending records might also be driven by the fact that, while social media profiles represent a single individual, spending is not necessarily an expression of an individual's personal preferences, as a large proportion of spending is both on fixed costs (e.g., groceries, bills), as well as spending on others (e.g., partners, children). This may add additional noise to the transaction records, lowering the predictive accuracy of the spending features in comparison to other types of behavioral footprints.

DISCUSSION

Our findings contribute to research on the automatic prediction of psychological traits by testing whether digital records of spending can be used to predict personality at scale. While our current predictions include considerable error, they provide greater predictive accuracy for more focused psychological traits (materialism and self-control) and were generally robust across participants with different levels of financial resources, varying slightly based on the deprivation levels of participant's local area on some traits.

Consistent with the theoretical distinction between "behavioral residues" and "identity claims"; predictive accuracies from spending records were modest compared with past research using social media data (e.g. Youyou et al., 2014). However, given these differences may also be due to sample composition, size, measurement error in the survey, the cultural characteristics of respondents, or other variables, there is a need for further research to test between these behavioral footprints directly by collecting multiple digital records (e.g. Facebook Likes and spending records) from the same sample of participants.

Predicting personality from spending also raises serious challenges for ethics and privacy. Firms could use personality predictions to identify and target vulnerable individuals, such as those low in self-control, with persuasive advertising for products harmful to their welfare, as well as harmful to society at large (e.g., adverts for gambling or smoking). Because personality predictions generated through one domain, such as spending records, can be used to target those same individuals in others, such as through direct mail, it is increasingly difficult for individuals to escape this new form of automatic psychological assessment and its downstream applications, should they wish to do so. This means that as personality predictions become more accurate and ubiquitous, and behavior is recorded digitally at an increasing scale, there is an urgent need for policymakers to ensure that individuals (and societies) are protected against potential abuse of such technologies.

References

Belk, R. W. (1985). Materialism: Trait Aspects of Living in the Material World. *Journal of Consumer Research*, *12*(3), 265. https://doi.org/10.1086/208515

Costa, P., & McCrae, R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*(1), 5–13. https://doi.org/10.1037//1040-3590.4.1.5

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. In *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011* (pp. 149–156). https://doi.org/10.1109/PASSAT/SocialCom.2011.33

Gosling, S. D., Ko, S., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398. https://doi.org/10.1037//0022-3514.82.3.379

Howell, R. T., Pchelin, P., & Iyer, R. (2012). The preference for experiences over possessions: Measurement and construct validation of the Experiential Buying Tendency Scale. *Journal of Positive Psychology*, *7*(1), 57–71. https://doi.org/10.1080/17439760.2011.626791

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5. https://doi.org/10.1073/pnas.1218772110

Matz, S. C., Gladstone, J. J., & Stillwell, D. (2016). Money Buys Happiness When Spending Fits Our Personality . *Psychological Science*, (1999), 1–15. https://doi.org/10.1177/0956797616635200

Mojtabai, R. (2013). Clinician-identified depression in community settings: Concordance with structured-interview diagnoses. *Psychotherapy and Psychosomatics*. https://doi.org/10.1159/000345968

Mosca, I., & McCrory, C. (2016). Personality and wealth accumulation among older couples: Do dispositional characteristics pay dividends? *Journal of Economic Psychology*, *56*, 1–19. https://doi.org/http://authors.elsevier.com/a/1T3nxc5GS77QJ

Nave, G., Minxha, J., Greenberg, D. M., Kosinski, M., Stillwell, D., & Rentfrow, J. (2018). Musical Preferences Predict Personality: Evidence From Active Listening and Facebook Likes. *Psychological Science*. https://doi.org/10.1177/0956797618761659

Oaten, M., & Cheng, K. (2007). Improvements in self-control from financial monitoring. *Journal of Economic Psychology*, *28*(4), 487–501. https://doi.org/10.1016/j.joep.2006.11.003

Park, G., Schwartz, H. A., Eichstaedt, J., Kern, M. L., Kosinski, M., Stillwell, D., … Seligman, M. E. P. (2014). Automatic Personality Assessment Through Social Media Language. *Journal of Personality and Social Psychology*, *108*(6), 934–952.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in*

Personality, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Richins, M. L., & Dawson, S. (1992). A consumer values orientation for materialism and its measurement: Scale development and validation. *Journal of Consumer Research*, *19*(3), 303–316. https://doi.org/10.1086/209304

Segalin, C., Perina, A., Cristani, M., & Vinciarelli, A. (2017). The Pictures We Like Are Our Image: Continuous Mapping of Favorite Pictures into Self-Assessed and Attributed Personality Traits. *IEEE Transactions on Affective Computing*, *8*(2), 268–285. https://doi.org/10.1109/TAFFC.2016.2516994

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Youyou, W., Kosinski, M., & Stillwell, D. (2014). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, *112*(4). https://doi.org/10.1073/pnas.1418680112