NORTHWESTERN UNIVERSITY

Beyond Traditional Measures of Personality with BISCUIT and BARE: A New Statistical Learning Technique and Behavioral Item Pool to Push Personality Psychology Forward

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By

Lorien Grey Elleman

EVANSTON, ILLINOIS

September 2020

Abstract

This dissertation investigates two ways in which personality psychology should move beyond the traditional approach of measuring personality with broad domains composed of trait descriptors, as exemplified by the Big Five taxonomy. The first study (Chapter 2) suggests an alternative to the traditional approach of aggregating personality items into domains. Mounting evidence indicates that, compared to domains, narrower measures of personality account for more variance in criteria and describe personality-criterion relationships more accurately. Analysis of individual personality items is the most granular approach to studying personality and is typically performed with statistical learning techniques (SLTs). The first study: (a) champions a new statistical learning technique, BISCUIT; (b) finds that BISCUIT provides a balance between prediction and parsimony; and (c) replicates previous findings that the broadness of the Big Five traits hinder their predictive power.

The second study (Chapter 3) suggests an alternative to the traditional approach of measuring personality with trait descriptors, or "traditional personality items." Of the three patterns commonly associated with personality (cognitions, emotions, and behaviors), behaviors are the least studied; traditional personality items tend to measure cognitions and emotions. Historically, yearlong patterns of specific behaviors have been thought of as criteria of personality measures, but the second study posits they should be classified as personality items because they measure patterns of behavior, a component of personality. The second study reviews and extends two pilot studies that indicated behavioral frequencies predict life outcomes, sometimes better than traditional personality items. The second study: (a) estimates the extent to which behavioral frequencies strengthen personality-criterion relationships above traditional personality items; (b) determines that some criteria are differentially predicted by personality item type; and (c) publishes an updated, public-domain item pool of behavioral frequencies: the BARE (Behavioral Acts, Revised and Expanded) Inventory.

Acknowledgements

I sincerely thank:

William Revelle, for being a patient, impulsive, generous, curious, wise, loyal, disagreeable, and kind mentor and friend. It has been a pleasure to work with someone who was so blunt when I was incorrect, so supportive when I was on the right track, and so excited when I suggested something he had not before considered.

Dan Mroczek, for your mentorship, encouragement, jovialness, insight, and generosity. I can't thank you enough.

Rick Zinbarg, for joining this committee on short notice and in the middle of a pandemic. I truly appreciate it.

David Condon, for returning my email seven years ago when I inquired about an RAship in the PMC lab. And for being a long-time mentor and collaborator. Most of my graduate school research has depended upon your commitment to SAPA.

Sarah McDougald, for our three-year collaboration. You taught me a great deal more than I had anticipated, and I am proud to have helped you pursue your own PhD.

My wonderful partner Marissa, for loving and supporting me in ways no one else ever has. And for copy editing.

Preface

I have structured this dissertation in a manner I have heard called the "European model." That is, each of Chapters 2 and 3 is an independent paper that has been or will be submitted for publication, but together they function as parts of a cohesive dissertation. Currently, Chapter 2 is *in press* in a special issue of the European Journal of Psychological Assessment, "New Approaches Towards Conceptualizing and Assessing Personality." I plan to submit Chapter 3 for publication immediately following my defense, and there should be a preprint available by the time this dissertation is widely available. If you would cite Chapters 2 or 3, please instead cite the published paper or preprint. You should be able to find links to them at https://www.researchgate.net/profile/Lorien_Elleman or https://lorienelleman.com/ for the next few years.

My reason for designing my dissertation this way was because, even before COVID-19 ravaged the economy, the academic job market was bleak. New psychology PhDs are being churned out of universities an order of magnitude faster than the rate at which tenured psychology professors are retiring. The idea that a dissertation should be an end in itself, and perhaps a publication as an afterthought, does not reflect the kind of competition that PhD graduates face. Publishing chapters from my dissertation before defending was a practical way to bolster my preparedness for entering the job market.

Table of Contents

List of Tables **8**

List of Figures **16**

1 General Introduction **18**

1.1 An Existential Limitation of the Big Five 20

1.2 A Practical Limitation of the Big Five 23

1.3 The Current Studies:
Moving Beyond the Tradition of the Big Five 27

2 That Takes the BISCUIT **28**

2.1 Abstract 28

2.2 Introduction 29

2.2.1 The Four Statistical Learning Techniques to be Compared 31

2.2.2 Aims of the Study 34

2.3 Methods 35

2.3.1 Sample 35

2.3.2 Measures 35

2.3.3 Procedure 36

2.3.4 Statistical Analyses 37

2.4 Results . . . 39
2.4.1 Predictive Accuracy . . . 39
2.4.2 Parsimony . . . 42
2.4.3 Post-hoc Analysis . . . 42
2.5 Discussion . . . 44
2.5.1 Limitations of the Study . . . 47
2.5.2 Future Directions . . . 48
2.5.3 Conclusions . . . 49

3 Laying Personality BARE **51**
3.1 Abstract . . . 51
3.2 Introduction . . . 52
3.2.1 Two Pilot Studies Have Examined the Incremental Validity of Behavioral Frequencies . . . 54
3.2.2 Overview of the Current Study . . . 56
3.3 Methods . . . 57
3.3.1 Participants . . . 57
3.3.2 Measures . . . 58
3.3.3 Statistical Analyses . . . 60
3.4 Results . . . 62
3.4.1 The Strength of Personality-Criterion Relationships Using Different Item Pools . . . 62
3.4.2 The Types of Personality Items Most Related to Each Criterion . . . 63
3.4.3 Summary of Best Items Content for Each Criterion . . . 64
3.5 Discussion . . . 70
3.5.1 Limitations . . . 73

3.5.2 Future Directions 74
3.5.3 Conclusions 75

4 General Discussion **77**

References **81**

Appendix A Appendix for Chapter 2 **100**

Appendix B Appendix for Chapter 3 **113**

List of Tables

3.1 Tests to determine the differences in non-independent correlations of criteria with BISCUIT models that use traditional personality item pools, compared to other item pools. Each test determines if correlation r_{AB} is significantly different from r_{AC}, accounting for r_{BC}. Variables A are six criteria. Variables B are BISCUIT models built using the traditional personality item pool. Variables C are BISCUIT models built using either the behavioral frequency item pool or an item pool that combined both personality item types. Bolded p-values indicate significant differences ($p < .05$), and bolded item pools and correlations indicate the models with the larger correlations. 64

3.2 Pearson's chi-squared tests comparing the frequencies of behavioral and traditional items in the total item pool against the frequencies in each BISCUIT model that was built with the total item pool. A bolded row indicates statistical significance ($p < .05$). 65

3.3 The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with general health ($R = .51$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 66

3.4 The 20 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with overall stress ($R = .52$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 66

3.5 The 14 personality items most strongly correlated with **body mass index**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with BMI ($R = .48$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 67

3.6 The 10 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with smoking frequency ($R = .53$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 68

3.7 The 10 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .51$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 69

3.8 The 10 personality items most strongly correlated with **emergency room visits**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a moderate correlation with emergency room visits ($R = .29$). Each listed correlation is an average across ten folds, and its standard error $\approx$ .02. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet. 69

A.1 Descriptive statistics of participant variables, for the initial sample (no restrictions) and the final sample (complete data for personality items and criteria). 101

A.2 Average pairwise administrations for training and test subsamples, by item pair type, for all data missingness conditions. Standard deviations are in parentheses. Average item-to-criterion pairwise administrations are larger than item-to-item administrations because criterion data were complete in all conditions. 101

A.3 Number of items selected by **BISCUIT,** for each criterion and level of personality data missingness. 101

A.4 Number of items selected by **the lasso,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations. 102

A.5 Number of items selected by **the elastic net,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations. 102

A.6 The six-item model selected by the BISCUIT to predict **Sleep Quality in the complete data condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion. 102

A.7 The one-item model selected by the BISCUIT to predict **BMI in the 90% data missingness condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion. 104

A.8 The one-item model selected by the BISCUIT to predict **General Health in the 90% data missingness condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion. . 104

A.9 Number of items selected by **BISCUIT run on imputed data,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations. Mean number of items = 39; median = 37. 104

A.10 Predictive accuracy (measured in multiple R) of **the lasso and elastic net,** based on personality data, across five levels of imposed missingness of data, in five criteria. 106

A.11 Predictive accuracy (measured in multiple R) of **BISCUIT and the random forest,** based on personality data, across five levels of imposed missingness of data, in five criteria. 107

A.12 Predictive accuracy (measured in multiple R) of **regression using the SPI-27 and regression using the Big Five,** based on personality data, across five levels of imposed missingness of data, in five criteria. 108

A.13 Predictive accuracy (measured in multiple R) of **BISCUIT using weighted coefficients and BISCUIT using imputed data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria. 109

A.14 Predictive accuracy (measured in multiple R) of **the elastic net and BISCUIT applied to the SPI-27 (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria. 110

A.15 Predictive accuracy (measured in multiple R) of **the elastic net and SPI-27 trained on imputed data and tested on complete data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria. 111

A.16 Predictive accuracy (measured in multiple R) of **BISCUIT trained on missing data and BISCUIT trained on imputed data, both tested on complete data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria. 112

B.1 Items of the BARE Inventory. "SAPA ID" refers to the unique item identifier used by the SAPA Project. "Origin" refers to whether the item was taken from the ORAIS or the BAI. The last column lists either the ORAIS facet associated with the item or the original BAI item ID number. 116

B.2 Personality items removed from BISCUIT's analysis of best items for a criterion due to the item content being synonymous with the criterion. Item correlation with the corresponding criterion is listed. 130

B.3 Reliabilities of BISCUIT models as if they were personality scales, by criterion. Items used per model and correlation with appropriate criterion are also listed. 131

B.4 The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with general health ($R = .48$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 132

B.5 The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with general health ($R = .46$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 132

B.6 The 20 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with overall stress ($R = .52$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 133

B.7 The 10 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a moderate correlation with overall stress ($R = .35$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 133

B.8 The 41 personality items most strongly correlated with **BMI**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with BMI ($R = .42$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 134

B.9 The 10 personality items most strongly correlated with **BMI**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with BMI ($R = .44$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. . . . 134

B.10 The 26 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a moderate correlation with smoking frequency ($R = .29$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 135

B.11 The 10 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with smoking frequency ($R = .54$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 135

B.12 The 27 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .41$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 136

B.13 The 10 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .49$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 136

B.14 The 10 personality items most strongly correlated with **ER visits**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a moderate correlation with ER visits ($R = .19$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 137

B.15 The 10 personality items most strongly correlated with **ER visits**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a moderate correlation with ER visits ($R = .24$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. 137

List of Figures

2.1 A visual representation of the three steps in which the sample data were prepared for analyses. (a) The final sample (complete data) was randomly split into the training sample (75% of the sample) and the test sample (25% of the sample). (b) For both the training and test samples, new data sets were created in which random missingness was imposed in the personality data. This representation only shows a data set in which 50% missingness was imposed, but this procedure was also performed for 25%, 75%, and 90% missingness. (c) For each data set with missing personality data, a new data set was created in which the missing data were imputed. For levels of missingness in which multiple imputation was used, twenty data sets were created for each data set with missing data. 36

2.2 Predictive accuracy (measured in multiple R) of the six statistical techniques, using personality data, across five levels of imposed data missingness, in five criteria. 41

2.3 Percentage reduction in predictive accuracy (R^2) for each of three techniques, averaged across five criteria. Each model was trained on one of five levels of imposed data missingness and tested on complete data. 43

3.1 Correlation of BISCUIT models with six criteria, using three pools of personality items (traditional items, behavioral frequencies, and a combined pool). The height of each shape is approximately the size of the estimate's 95% confidence interval. 62

A.1 Predictive accuracy (measured in R^2) of the six statistical techniques, using personality data, across five levels of imposed data missingness, in five criteria. 103

A.2 Predictive accuracy (measured in multiple R) of four techniques based on personality data, across five levels of imposed missingness of data, in five criteria. **Models were only trained on data with an imposed data missingness level; the predictive accuracy of each technique was tested on complete data.** 105

Chapter 1

General Introduction

> When we propound a general theory in our sciences, we are sure only that, literally speaking, all such theories are false. They are only partial and provisional truths which are necessary to us, as steps on which we rest, so as to go on with investigation; they embody only the present state of our knowledge, and consequently they must change with the growth of science, and all the more often when sciences are less advanced in their evolution.
>
> —Claude Bernard, *An Introduction to the Study of Experimental Medicine*

Human personality is broadly conceived of as individual differences in patterns of thinking, feeling, and behaving. The "lexical hypothesis" is an approach that describes these patterns as trait descriptors (i.e., adjectives, short phrases, or sentences) by leveraging the assumption that important individual differences will appear as descriptions in human languages (Allport and Odbert, 1936; Goldberg, 1990). The "Big Five" taxonomy of personality categorizes trait descriptors into five broad traits: Conscientiousness, Agreeableness, Neuroticism, Openness to Experience, and Extraversion (Goldberg, 1990). In terms of number of studies, the Big Five taxonomy is undoubtedly the most dominant of trait taxonomies in personality psychology (John et al., 2008). It has been pervasive in the field for approximately two decades (Funder, 2001, p. 200), and its enduring popularity has cemented it as

the default measure of personality.

The Big Five's lasting dominance is due, in part, to the many conveniences of its taxonomy. One such convenience is the small number of traits one needs to remember. Findings from cognitive psychology suggest that individuals can hold only a small number of concepts in their short-term memory (7 ± 2, Miller, 1956; or, more recently, 4 ± 1, Cowan, 2001). In addition to being in this cognitive sweet spot, the Big Five describe personality at "the broadest level at which categories still have a high degree of fidelity" which is the level at which people prefer to describe others and themselves (John et al., 1991, p. 348). Since Big Five scales, like most measures of personality, are typically self-reported (Baumeister et al., 2007), administering them is also convenient. And conveniently, there are scales of the Big Five that may fit any researcher's needs: scales that have hundreds of items (e.g., Costa and McCrae, 1992; Condon, 2014), dozens (e.g., John and Srivastava, 1999), or even less than a dozen (e.g., Gosling et al., 2003); established scales that require costly licenses (e.g., Costa and McCrae, 1992), or free-to-use measures that are psychometrically sound (e.g., Goldberg et al., 2006; Condon, 2014).

The Big Five's popularity has also helped it become even more popular, as it has provided a set of common domains for researchers. In the middle of the 20th century, the number of personality constructs proliferated such that simply documenting them was a significant undertaking (Goldberg, 1971). The Big Five have given psychologists a set of personality traits that function as a common language amidst a "Babel of concepts and scales" (John and Srivastava, 1999, p. 102), and have arguably shaped personality psychology into the "unified scientific discipline" that Eysenck (1991) called for (p. 786). Broader and narrower personality traits have been integrated hierarchically above and below the Big Five. Broader traits are super-factors of the Big Five (John et al., 2008), while narrower traits are considered to be "facets" (Costa and McCrae, 1992), "aspects" (DeYoung et al., 2007), or "nuances" (McCrae, 2015) of the Big Five.

Of course, the Big Five are popular for more than just their convenience. There is a long history of the Big Five structure being found as the optimal solution in factor analytic studies, dating back to the 1940's (Fiske, 1949; the Big Five were found across three analyses, but not within any one of them). Since then, dozens of other studies of English-language assessments have found a factor structure similar to the Big Five (for a review, see John et al., 2008). In one of the largest studies of its kind, Goldberg (1990) found the Big Five factors among hundreds of descriptors, across multiple samples. Much work has also been done in non-English samples, and it appears that the Big Five are relatively universal; the factor structure has been found across dozens of cultures (for a review, see Allik et al., 2012).

Lastly, the Big Five continue to be popular because studies continue to find that they are related to phenomena of psychologists' interests. For example, a vast body of evidence has linked the Big Five with important life outcomes (see Ozer and Benet-Martínez, 2006 and Roberts et al., 2007 for reviews). Additionally, the Big Five have been found to be associated with biology and behavior; the Big Five are heritable (e.g., Loehlin, 1998), differentially related to the volume of brain regions (e.g., DeYoung et al., 2010), and associated with actual behavior (Grucza and Goldberg, 2007; Jackson et al., 2010). It is likely that the Big Five will remain popular until researchers can no longer milk findings from the taxonomy.

1.1 An Existential Limitation of the Big Five

An existential limitation of the Big Five is that they may not actually exist; they are perhaps more accurately identified as convenient fictions (Revelle, 1983). There is a long-standing debate concerning whether the Big Five, or any latent traits, are "real" and/or have causal power (see Mõttus, 2016 and Asendorpf et al., 2016 for a recent example). However, two facts are indisputable: one, the Big Five were constructed with factor analysis; and two, factor analysis is merely a mathematical simplification of data. Researchers have no hope

of comprehending all at once the hundreds, thousands, or millions of cells in a table that represent the data of a psychological study. Thus, we sometimes use the crutch of factor analysis and other models in order to simplify data. In the case of the Big Five, hundreds of descriptors are simplified into five broad ideas (e.g., Costa and McCrae, 1992), which we may more easily juggle in our minds (Cowan, 2001).

Factor analysis does not magically summon latent traits into physical existence or unveil a hidden absolute truth. The Big Five taxonomy is, at its core, an abstraction, and much like a true score it is "a Platonic ideal, a concept that exists in a non-spatial/non-temporal universe, in a world of pure forms somewhere out in the clouds" (Hogan and Foster, 2016, p. 4). In a manner of speaking, factor analysis actually leads away from the truth, insofar as it reduces complexity at the cost of information—information that is necessary to fully describe some phenomenon. Hogan and Foster (2017) provide a simple example of this loss of information by facetiously suggesting that psychologists should combine measures of height and weight into a factor called "size" because the correlation between the two is of a similar magnitude to that of intercorrelations between same-domain facets (p. 23).

The repeated, cross-cultural replication of the same factor solution does not indicate that the Big Five are real or the one true answer for personality. The most we may reasonably extrapolate from the cross-cultural replication of the Big Five is the following: If one were to vastly simplify the underlying covariance matrix of a broad pool of trait descriptors, the Big Five structure would probably be a decent solution to those data. In the same vein, evidence that the Big Five are heritable or related to biological mechanisms is not evidence that *only* the Big Five are associated with these phenomena. In fact, these sorts of findings from Big Five studies may generalize to any reasonable measure of personality; for example, personality appears to be heritable at every measurable breadth, from traits broader than the Big Five (Vukasović and Bratko, 2015) to individual personality items (Mõttus et al., 2017).

On the subject of heritability, it is worthwhile to take a short diversion into behavior genetics, as the evidence in that field suggests the opposite of the top-down causal chain that proponents of the Big Five taxonomy sometimes presume—that a few broad traits cause narrower traits, which then cause ever-narrower traits. While the first law of behavior genetics establishes the heritability of all behavioral traits (Turkheimer, 2000), the fourth law states that any one these traits "is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability" (Chabris et al., 2015, p. 305). It appears that the gene, a tiny unit of biological individual differences, is responsible for broader patterns of personality. But Meaney (2010) asserts this causal chain is more complicated:

> The operation of the genome at any phase of the life cycle is an emergent property of the constant and very physical interaction of the genome with environmentally regulated, intracellular signals that directly alter chromatin structure. Thus, function at any level of biology emerges as a function of the continuous dialogue between the genome and its environment. (p. 69)

When sufficiently aggregated, this constant dialogue is observable as narrow patterns of thinking, feeling, and behaving. Thus, rather than being Aristotelian *first uncaused causes* of personality, the Big Five are instead executive summaries of gene-by-environment interactions.

That the Big Five may not actually exist or have causal power is not necessarily a problem for research in personality psychology. Decades of studies concerning the Big Five have helped researchers comprehend otherwise incomprehensible data and discover associations between personality and many important phenomena. So long as Big Five measures prove adequate in the continuing advancement of psychological research, there would be no reason to move beyond analysis of the Big Five.

1.2 A Practical Limitation of the Big Five

A practical limitation of the Big Five is that they no longer prove adequate in the continuing advancement of psychological research. Specifically, the broadness of the Big Five traits both diminishes the explanatory power of personality and limits the specificity of personality-criterion relationships. Narrower personality traits, namely, facets and items, have proven to be useful alternatives to the Big Five.[1]

Facet-sized traits are more predictive and informative than the Big Five.

The term "facet" was popularized by Costa and McCrae (1985) and denoted six subtraits within each of Neuroticism, Extraversion, and Openness to Experience. A revised version of the NEO Personality Inventory (NEO-PI-R; Costa and McCrae, 1992) added six facets for each of Agreeableness and Conscientiousness. The fact that six facets were placed beneath each Big Five trait was not decided on strictly psychometric grounds and possibly was for tidiness, convenience, or marketing. Facets of the NEO Big Five are "observable characteristics of the individual that go beyond the five factors" (Costa and McCrae, 2008); the specific variance of the facets are stable (Mõttus et al., 2014) and heritable (Jang et al., 1998). Other researchers have followed suit in describing lower-level subtraits of broader domains as "facets" for other inventories based on the Big Five (Hofstee et al., 1992) or inventories of other taxonomies (e.g., HEXACO; Lee and Ashton, 2004).

For almost as long as the Big Five have been ubiquitous, research has indicated that narrower traits account for more variance in criteria than broader domains. Early studies did not actually compare facets with the Big Five, but instead examined the difference between broader and narrower traits (while correcting for shrinkage). For example, Mershon

[1] DeYoung et al. (2007) split each Big Five trait in two, with each pair forming "aspects" of a domain. Aspects are broader traits than facets but narrower than the Big Five. The difference between aspects and facets will not be explored. Occasionally, I refer to aspect-level research in order to compare domains with lower-level personality traits.

and Gorsuch (1988) found that, compared to a six-trait inventory, a 16-trait inventory of personality traits accounted for more variance in 15 of 17 "real-life" criteria; and Paunonen (1998) found that, across two studies, more incremental variance in criteria was found for more numerous and narrow traits (22 and 16) over and above the Big Five, than vice versa.[2]

It appears that the first example of comparing the explanatory power of a Big Five measure with its facets was from Paunonen and Ashton (2001). In this study, variance explained in 40 criteria was compared for each of two measures of the Big Five (the PRF-JPI and the NEO-PI-R) against their corresponding facets (numbering 34 and 30, respectively.) Interestingly, instead of comparing each set of Big Five scales against all of its facets, Paunonen and Ashton only used five facets per criteria, and these five facets were selected in advance from a panel of expert judges. Even with this limitation of five facets per criterion, sets of five facets tended to significantly predict more criteria and more average variance in criteria than their corresponding Big Five scales. In a follow-up extension, Paunonen et al. (2003) found that the relationships between personality and 19 criteria were more consistently replicated across four cultures for 10 facets of the Supernumerary Personality Inventory (SPI)[3] than for either of two domain measures (the Big Five NEO-FFI and the three-factor SPI).

Facets or aspects of the same domain can differentially correlate with a criterion, and when they do, these lower-level traits are better at specifying personality-criterion relationships than the broader domains. For instance, Mõttus et al. (2012) found that diagnosis history of a sexually transmitted disease was related to just three facets (Deliberation, Hostility, and Impulsivity), as opposed to two Big Five domains (Agreeableness and Neuroticism). In a geographical psychology study, Rentfrow (2014) observed an instance of two aspects

[2]I refer to such narrow traits as "facet-sized" traits to denote their narrow scope, even though they are not components of a larger domain. One example of why this is useful can be observed in the 27-factor personality inventory of Condon (SPI-27; 2018): The SPI-27 are not subtraits of broader domains, but are "facet-sized" due to their being as narrow in scope as the NEO-PI-R facets (e.g., "Order" and "Anxiety" are both traits of the SPI-27 and facets of the NEO-PI-R, and they are of similar breadth in both inventories).

[3]This measure is unrelated to the SPI-27 of Condon (2018).

of Extraversion (Assertiveness and Activity) differentially correlating with health and social capital, such that signs were opposite for the two aspects. In another study of geographical psychology, Elleman et al. (2020) found that the population density and income disparity of ZIP Codes were related to the aggregated personality of ZIP Codes in three domains of the Big Five (Openness, Conscientiousness, and Agreeableness), but were primarily related to just six and seven (respectively) of the possible 18 facets of those three domains. In a frequently-cited health psychology study, body mass index (BMI) was positively correlated with the facet Impulsivity, but negatively related to its domain, Neuroticism (Terracciano et al., 2009). Lastly, a meta-analysis of the higher-order trait Grit found that its facet Perseverance of Effort was a better predictor of academic achievement than overall Grit or the other facet of Grit, Consistency of Interest (Credé et al., 2017). In all of these examples, facet-level analysis further specified the personality-criterion relationships beyond what the domains could have.

Items are more predictive and informative than the Big Five and facets.

In personality psychology, the term "nuance" was originally used by McCrae (2015) to denote personality traits below the facet level, which had unique item variance not accounted for by higher-order traits and were "the absolute bottom of the trait hierarchy" (p. 99). Functionally, nuances are the individual items in a given pool of personality items, whether or not higher-order traits are presupposed in that pool. In practice, studies have tended to conceptualize nuances as items that are part of a higher-order scale, and contrast nuances with those higher-order traits. An initial wave of evidence in the last decade indicates that nuances are both reliable and valid measures of personality; nuances have longitudinal rank-order stability, have cross-rater agreement, and like all stable measures of personality, are heritable (Mõttus et al., 2014, 2017, 2019).

Traditional approaches to personality scale construction (e.g., Loevinger, 1957) would

consider personality items to be both "samples" and "signs." That is, personality items are nothing more than sample behaviors which point to broader underlying traits. In that view, the unique variance associated with items would not be of particular importance; in the same way that factor analysis simplifies data by omitting information, a scale of a higher-order trait omits the unique variance of its component items. This omission ultimately impedes the field of psychometrics, whose task is "to isolate, to identify and, so far as possible, to measure separately the important components of variance" (Loevinger, 1957, p. 649).

Historically, psychometricians have not considered item-level variance to be an important component of variance due to limitations of sample sizes. Goldberg (1993), however, had the foresight to imagine samples so large that "it would be silly even to amalgamate the items into scales because one would inevitably lose some specific variance at the item level that could serve to increase predictive accuracy" (pp. 181-182). Growing evidence over the past few years seems to confirm Goldberg's claim: in several studies with large samples, personality-criterion models built with items were more predictive than those built with facets or domains (Seeboth and Mõttus, 2018; Mõttus et al., 2015, 2020).

In studies in which the predictive power of items was examined individually, instead of in aggregate, item-level findings provided the greatest amount of information in describing personality-criterion relationships. For instance, in the previous example in which BMI was positively related to Impulsivity, Terracciano et al. (2009) also found that BMI was associated with only two Impulsivity items, both of which measured overeating. In another previous example, ZIP Code population density was linked to Dutifulness, Morality, and Orderliness, but an examination of items indicated a more precise finding: that each facet was associated with population density because it contained one or two items related to having an aversion to rules (Elleman et al., 2020). Item-level results indicated a tight cluster of items (perhaps broad enough to be considered a facet) whose items were dispersed across three facets.

Due to the limited number of studies that have investigated personality at the item level,

the full utility of this type of research is largely unknown. Researchers seem to agree that personality items generally account for more variance in criteria than domains or facets, but what is less known is whether one should expect items to be more informative, above and beyond domains and facets, in describing personality-criterion relationships. There are few instances in the personality literature of a relationship between a criterion and a high-order personality trait being explained by a few items, but this dearth appears to be due to a lack of inquiry. Even in the case of BMI and Impulsivity, most follow-up studies of Terracciano et al. (2009) reported only the facet-level relationship and did not examine whether the items of Impulsivity differentially correlated with BMI (Vainik et al., 2015).

1.3 The Current Studies: Moving Beyond the Tradition of the Big Five

In this introduction, I have made the case that measuring personality using broad summaries of trait descriptors, as exemplified by the Big Five, has existential and practical limitations. The two following studies both move beyond this traditional approach by investigating item-level relationships. The first study (Chapter 2) compares the predictive validity of the Big Five, 27 facet-sized traits, and four statistical learning techniques which use item-level analysis, including BISCUIT, a novel technique designed to analyze personality data. The second study (Chapter 3) posits that self-reported yearlong behavioral frequencies are measures of personality, and examines their incremental validity over and above traditional trait descriptors.

Chapter 2

That Takes the BISCUIT

Predictive Accuracy and Parsimony of Four Statistical Learning Techniques in Personality Data, with Data Missingness Conditions

2.1 Abstract

The predictive accuracy of personality-criterion regression models may be improved with statistical learning (SL) techniques. This study introduced a novel SL technique, BISCUIT (Best Items Scale that is Cross-validated, Unit-weighted, Informative and Transparent). The predictive accuracy and parsimony of BISCUIT was compared with three established SL techniques (the lasso, elastic net, and random forest) and regression using two sets of scales for five criteria across five levels of data missingness. BISCUIT's predictive accuracy was competitive with other SL techniques at higher levels of data missingness. BISCUIT most frequently produced the most parsimonious SL model. In terms of predictive accuracy,

the elastic net and lasso dominated other techniques in the complete data condition and in conditions with up to 50% data missingness. Regression using 27 narrow traits was an intermediate choice for predictive accuracy. For most criteria and levels of data missingness, regression using the Big Five had the worst predictive accuracy. Overall, loss in predictive accuracy due to data missingness was modest, even at 90% data missingness. Findings suggest that personality researchers should consider incorporating planned data missingness and SL techniques into their designs and analyses.

2.2 Introduction

Research over the last decade has indicated that personality items (often called "nuances;" McCrae, 2015) are both reliable and valid measures of personality. There is cross-rater agreement associated with the specific variance of nuances (Mõttus et al., 2014) and nuances have rank-order stability over time, and are heritable (Mõttus et al., 2017, 2019). Additionally, personality-criterion models that utilize nuances tend to be more predictive than those that employ broad domains (e.g., the Big Five; Goldberg, 1990) or narrower facets (Seeboth and Mõttus, 2018; Mõttus et al., 2015, 2020).

Item-level analysis requires a number of multiple comparisons that is an order of magnitude greater than broad personality domains or narrower facets. Traditional methods of analysis, such as regression, can overfit the data or find few stable results after statistical adjustments. Recently, several researchers have suggested using statistical learning (SL) techniques[1] to study nuances (Chapman et al., 2016) and improve the prediction of outcomes in personality psychology (Yarkoni and Westfall, 2017). Compared to traditional statistical methods, many SL techniques are more complex and better suited to the study of nuances

[1]Specifically, supervised learning. Models generated by supervised learning techniques are "supervised" by the criterion variable they predict. Unsupervised learning techniques describe patterns in data without the use of a criterion.

because they have been designed to reduce overfitting. Usually, the accuracy of an SL model is measured by the prediction of a holdout sample (the "test sample") that has been kept separate from the sample upon which the model was built (the "training sample"). For an overview of statistical learning, see James et al. (2017); for short overviews, see Chapman et al. (2016) and Yarkoni and Westfall (2017).

To improve prediction of the test sample, an SL technique may augment a basic statistical method, such as regression, in several ways. For instance, an SL technique may implement "regularization" to shrink the coefficients of a model to reduce overfitting (e.g., ridge regression; Hoerl and Kennard, 1970). Some SL techniques use "variable selection" to retain the most important variables for the final model (e.g., the lasso; Tibshirani, 1994). SL techniques may test many different models via "resampling," an iterative sampling procedure: each new model is developed iteratively on randomly selected sub-samples of the training data and may be cross-validated using holdout portions of the training data (for a review of using cross-validation for model selection, see Arlot and Celisse, 2010). Resampling procedures may be used to aggregate the different models into a final model, to estimate the error of the model estimates, and/or to optimize model hyperparameters (or "tuning parameters"). A tuning parameter differs from a typical model parameter in that the researcher preselects a series of tuning parameter coefficients. Each tuning parameter coefficient is input into a new model or series of models. Hyperparameters are tuned (i.e., an optimal value is found for each) by selecting the model or aggregated model with the lowest cross-validated error. For example, the lasso's regularization hyperparameter must be tuned in order to determine the optimal degree of regularization for a particular criterion (Tibshirani, 1994).

Applying certain SL techniques to personality psychology may result in final models that are substantially more complex, and perhaps more difficult to interpret, than traditional personality models. For example, in applying an SL technique to personality data, Seeboth and Mõttus (2018) took an approach that was similar to a genome-wide association study

(GWAS; Hirschhorn and Daly, 2005), such that personality-criterion associations were considered to be "driven by a large number of specific personality characteristics" (p. 188) and nuance-criterion relationships were summarized by the variance explained by using an unspecified number of items. Even if nuances predict a criterion better than facets or domains, certain SL methods, such as a "persome"-wide association study (Mõttus et al., 2020), may output a model with as many or nearly as many predictors as there are items in the pool. While predictive accuracy and parsimony differ for each SL approach, very little, if any, research in personality psychology has been performed to compare the predictive accuracy and parsimony of SL techniques.

2.2.1 The Four Statistical Learning Techniques to be Compared

BISCUIT. The Best Items Scale that is Cross-validated, Unit-weighted, Informative and Transparent, or BISCUIT (Revelle, 2020), is a correlation-based SL technique that grew out of the practical need for generating parsimonious models to describe nuance-level relationships in Massively Missing Completely At Random (MMCAR) data (Revelle et al., 2010, 2016).[2] Similar to the "criterion-keyed scale construction" of Chapman et al. (2016) and reminiscent of the procedures used in the development of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway and McKinley, 1942), BISCUIT utilizes variable selection to retain the items that most strongly correlate with a criterion (i.e., the best items). Item-level correlations in BISCUIT are calculated solely from pairwise administrations of items. Thus, unlike other SL techniques in this study, BISCUIT may be run on MMCAR data structures without the need for imputation. BISCUIT uses a resampling procedure to determine a cross-validated list of the best items based upon the average correlation; either bootstrap

[2]In MMCAR data, each participant is given a random sample of items; the raw data are mostly (i.e., massively) missing, but this missingness has been completely randomized. Individual scales may be over- or undersampled.

aggregation ("bagging") or k-fold cross-validation may be utilized (for a description of bagging, see Breiman, 1996; for k-fold cross-validation, see Chapman et al., 2016, p. 607). The cross-validated best items are combined into a scale for the criterion, which is the final model for BISCUIT. In BISCUIT's empirically constructed scale (and typical personality scales), all best items are weighted the same (i.e., unit-weighted).[3]

Compared to an optimally weighted regression model, a unit-weighted model tends to fit the initial data set about as well (e.g., Wilks, 1938; Dawes, 1979), and often has improved predictive accuracy in new data sets (Wainer, 1976; Waller, 2008); optimal weights are optimal only for the initial data set, and overfitted in others. Although there is only one set of optimal weights for a least-squares regression model, there are an infinite number of alternative sets of weights for a more robust, non-least-squares solution (Waller, 2008). BISCUIT employs unit-weighting as a simple alternative to least-squares regression for the same reason that regression-based statistical learning techniques implement regularization: to improve upon the predictive accuracy of an overfitted regression model by systematically modifying the model's coefficients. Lastly, BISCUIT's unit-weighted models and output are like oven windows through which one can view a biscuit baking; BISCUIT outputs a list of items that most highly correlate with a criterion, their correlations with the criterion, and the content of each item. BISCUIT's tuning parameter is the number of best items to select for a model.

To provide clarity around the BISCUIT algorithm, the following is a step-by-step procedure for it: (1) At least two options are selected: (1a) the range of N best items to be retained and (1b) whether the analysis should use bagging or k-fold cross-validation (this

[3]Reviewers were concerned that BISCUIT's performance would improve by weighting variables instead of unit-weighting them. An option to weight variables (equal to their zero-order correlations) has been added to the BISCUIT algorithm. Comparative analysis indicated that BISCWIT (Weighted, instead of Unweighted) performed sometimes better than BISCUIT, sometimes worse, and on average about the same (see Table A.13 in Appendix A). A reviewer commented that BISCWIT's performance could improve if its coefficients were estimated by multiple regression instead of zero-order correlation. We agree that exploring this modification in a future study would be worthwhile.

example will assume k-fold). (2) For a given criterion, for each of k splits: (2a) A criterion-by-item correlation matrix is calculated, based on the pairwise administrations of the raw data in the training subsample. (2b) The N items that have the largest correlations with the criterion are retained and formed into a unit-weighted scale. Both item-level and scale-level correlations are recorded. (2c) The holdout subsample may be used to determine the cross-validated correlation of the unit-weighted scale with the criterion. (3) The steps in 2 are repeated k times. (4) Average correlations across the k splits are found. (5) A final set of N items is retained, based on the number of items that were best cross-validated across the k splits. (6) The BISCUIT model is output as a scale, listing each item and whether it is negatively or positively associated with the criterion.

Lasso. The Least Absolute Shrinkage and Selection Operator, or lasso (Tibshirani, 1994), is a regression-based SL technique that was created to be an improvement over traditional regression and ridge regression (Hoerl and Kennard, 1970). The lasso and ridge regression are similar in that each uses a regularization penalty that is based on a tuning parameter and the magnitude of each regression coefficient. However, ridge regression's penalty (ℓ_2) uses the square of each coefficient, while the lasso's penalty (ℓ_1) uses the absolute value of each coefficient (see Equations A.1 and A.2 in Appendix A). The lasso's penalty, unlike ridge regression's penalty, allows regression coefficients to shrink to values of zero. After regularization, variables with zero-value coefficients are discarded, effectively giving the lasso a variable selection feature. The lasso's tuning parameter λ determines the magnitude of coefficient shrinkage.

Elastic Net. The elastic net is a regression-based SL technique that is framed as an improvement over the lasso (Zou and Hastie, 2005). The elastic net incorporates ridge regression and the lasso into one algorithm; the lasso is a special case of the elastic net when the λ_2 tuning parameter of the elastic net is set to 0, and ridge regression is a special case of the elastic net when λ_2 is set to 1 (Zou and Hastie, 2005). Two typical tuning parameters

of the elastic net are: (a) λ, which determines the magnitude of coefficient shrinkage; and (b) λ_2, which determines the extent to which groups of highly correlated variables will be retained.

Random Forest. The random forest (Breiman, 2001) is an SL technique based upon decision trees. A decision tree iteratively partitions a data set, one variable at a time, into two groups such that differences in the groups maximally predict a criterion. Essentially, the random forest combines the bagging resampling procedure with the random decision forest (Ho, 1995). In the random decision forest, a final model is built from an aggregation of multiple trees; in each tree, a random subset of predicting variables is selected for each branch. The random forest combines bagging and the random decision forest by aggregating bootstrapped decision tree models, where each model includes a subsample of predicting variables. The purpose of bagging and the random decision forest is similar: to aggregate models based upon samples from the available data in order to reduce overfitting. There are inconsistencies in the literature regarding what, if any, tuning parameters should be used for the random forest (Probst and Boulesteix, 2018; Tang et al., 2018).

2.2.2 Aims of the Study

The primary aim of this study was (a) using personality data, to compare the models of four SL techniques in terms of their predictive accuracy. Because of our particular interest in BISCUIT, and because BISCUIT was built to perform well with MMCAR data, we also evaluated (b) in terms of predictive accuracy, whether BISCUIT models gained an advantage over other SL models as the rate of data missingness was artificially increased in the sample. Finally, we determined (c) the extent to which BISCUIT tended to provide more parsimonious models than other SL techniques, which was quantified by the number of personality items used in a model.

2.3 Methods

2.3.1 Sample

Participant data were collected at SAPA-project.org, an international online personality assessment. The SAPA (Synthetic Aperture Personality Assessment) Project is an ongoing research project where each participant is given a small random sample of a large item pool (over 6,000 items), resulting in an MMCAR data structure. An initial sample of 497,048 participants (64% female; median age = 26 years; from 228 countries; 39% from the U.S.) was collected from February 7, 2017 to November 12, 2018. In order to run out-of-the-box algorithms for the lasso, elastic net, and random forest, the data were limited to complete cases for the selected personality items and criteria (see below in *Measures*). Requiring complete data reduced the sample to 78,828 participants. This final sample had participants who were from 200 countries (57% from the U.S.), 65% were female, and the median age was 33 years ($min = 14, max = 90$). Descriptive information concerning the initial and final samples are available in Table A.1 in Appendix A.

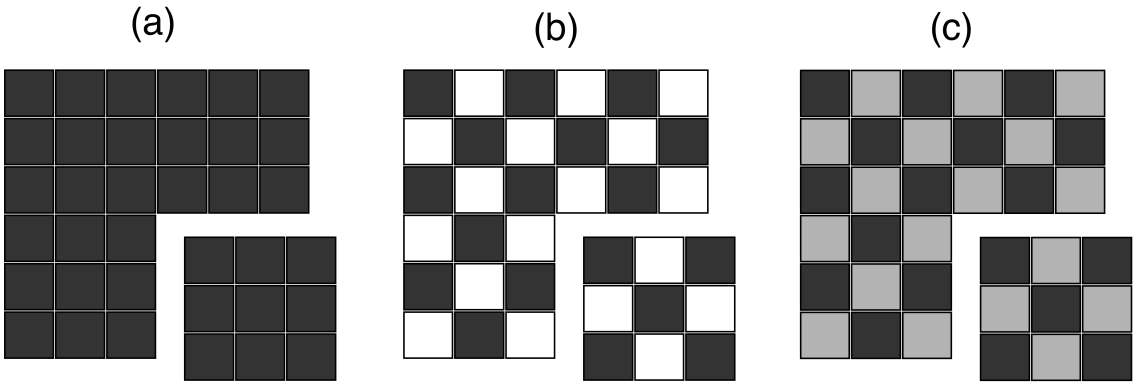
2.3.2 Measures

All measures were self-reported. Personality was measured with the 135-item SPI-27 (SAPA Personality Inventory; Condon, 2018), a personality inventory that may be scored as 27 traits (five items per trait) or as the Big Five domains (70 total items; 14 items per trait). Each personality item was answered on a six-point Likert-like scale. There were five criteria: Body Mass Index (BMI), smoking frequency, sleep quality, general health, and educational achievement. These specific criteria were selected for their breadth. Demographic measures included ethnicity (if the participant was from the U.S.), age, sex, and country of residence.

2.3.3 Procedure

All steps in the procedure and analyses were performed with the statistical programing language and environment R (R Core Team, 2019) in the integrated development environment RStudio (RStudio Team, 2019). There were three primary steps to preparing the data for analysis (Figure 2.1): (a) split the final sample into the test and training samples; (b) create new test and training sample data sets by imposing increasing levels of missingness; and (c) for each data set with missing data, create new data sets in which the missing data were imputed. More details of each step are described below.

Figure 2.1: A visual representation of the three steps in which the sample data were prepared for analyses. (a) The final sample (complete data) was randomly split into the training sample (75% of the sample) and the test sample (25% of the sample). (b) For both the training and test samples, new data sets were created in which random missingness was imposed in the personality data. This representation only shows a data set in which 50% missingness was imposed, but this procedure was also performed for 25%, 75%, and 90% missingness. (c) For each data set with missing personality data, a new data set was created in which the missing data were imputed. For levels of missingness in which multiple imputation was used, twenty data sets were created for each data set with missing data.



(a) The final sample was randomly split into the training sample (75% of participants) and test sample (the remaining 25%). Having the training sample be larger than the test

sample gives training models greater power and is typical (e.g., Breiman, 1996; Chapman et al., 2016; Seeboth and Mõttus, 2018).

(b) Because BISCUIT was designed to analyze MMCAR data, it was necessary to test whether missingness in personality data would give an advantage to BISCUIT's predictive accuracy over the models of other techniques. To do this, four new data sets were created (for each of the training and test samples), where each new data set imposed increasing levels of random missingness in the personality data (25%, 50%, 75% and 90% missingness; see Table A.2 in Appendix A for pairwise administrations at each level of data missingness).

(c) BISCUIT's algorithm can converge on data sets with missing data, but other out-of-the-box SL techniques cannot. Therefore, new data sets were created that imputed the imposed missing data (using the "MIPCA" and "imputePCA" functions of the R package "missMDA;" Josse and Husson, 2012, 2016). For data sets with 25%, 50%, and 75% data missingness, imputation was performed with multiple imputation using Bayesian principal components analysis (BayesMIPCA; Audigier et al., 2014). This imputation method performs favorably compared to other methods (Schmitt et al., 2015). However, BayesMIPCA did not converge on 90% data missingness, so a single imputation method that was similar to BayesMIPCA was used for 90% missingness data sets: single imputation using a regularized iterative principal components analysis (Audigier et al., 2016). For both imputation methods, the number of principal components was determined with parallel analysis (Horn, 1965).

2.3.4 Statistical Analyses

Analyses consisted of three steps: for each criterion and at each level of data missingness, (a) each model was built using the appropriate training data set; (b) using test personality data, each model predicted each criterion; and (c) the predictive accuracy of each model was

determined by calculating the multiple R value between a model's prediction of a criterion and the actual value of the criterion in the test data.[4] More details of each technique's procedures are described below.

BISCUIT. BISCUIT was run using the "bestScales" function in the "psych" package (Revelle, 2020, version 1.9.11) of R. BISCUIT was the only technique run on data sets with missing data. To increase the speed of computation, BISCUIT was set to use k-fold cross-validation ($k = 10$) instead of bagging. BISCUIT's tuning parameter, the number of best items, was given the full range of possible values, from one item to one hundred thirty-five items. An average model was found for each count of items, using k-fold cross-validation. Across counts of items, and for each criterion and level of missingness in the data, the model with the highest cross-validated multiple R was selected.

Lasso. The lasso was run using the "cv.glmnet" function in the "glmnet" package (Friedman et al., 2010) of R. The tuning parameter λ was optimized using the function's default sequence of values. An average model was found for each value of λ using k-fold cross-validation ($k = 10$). For each criterion and level of missingness in the data, the model with the lowest cross-validated error was selected.

Elastic Net. The elastic net was also run using the "cv.glmnet" function. For the tuning parameter λ_2, eleven values were tested, from 0 to 1 in increments of .1. For each value of λ_2, the tuning parameter λ was optimized using the function's default sequence of values. An average model was found for each value of λ_2 using k-fold cross-validation ($k = 10$). Across values of λ_2, and for each criterion and level of data missingness, the model with the lowest cross-validated error was selected.

Random Forest. The random forest was run using the "randomForest" function in the

[4]Multiple imputation generated twenty data sets for each level of data missingness. For each level of data missingness, twenty models were built using the twenty imputed training data sets, each model was applied to one of the twenty imputed test data sets, model fits were determined, and model fits were averaged across the twenty predictions.

"randomForest" package (Liaw and Wiener, 2002) of R. Forty-five personality items were sampled as candidates for each branch of each tree (which was the default value for the function). There were one hundred trees per forest model in order to maintain computational feasibility (i.e., less than one week of computation for all random forest models).

Regression. Two regression analyses were used as baselines for typical statistical analyses in personality psychology. One regression technique used the Big Five measures as predictors, while the other used the 27 traits of the SPI-27. These basic regression models did not implement any tuning parameters or resampling procedures. Given the high power of the study, all predicting variables were included in every regression model.

2.4 Results

2.4.1 Predictive Accuracy

Predictive accuracy of the techniques in 25 total conditions (five criteria by five levels of data missingness) was calculated with Multiple R and R^2 (R^2 was used to calculate ratios of predictive accuracy between models). The elastic net had the highest predictive accuracy in 13 conditions, BISCUIT in seven conditions, the lasso in three conditions, and regression using the SPI-27 in two conditions (Figure 2.2; Tables A.10 – A.12 in Appendix A. For R^2, see Figure A.1 in Appendix A). Additionally, the elastic net or lasso had the highest predictive accuracy for all five criteria for the complete, 25%, and 50% data missingness conditions. Models generated by the lasso were, on average, 99.8% as predictive as the elastic net models, which indicated that the predictive accuracy of the elastic net and lasso were functionally equivalent. For complete data, multiple R effect sizes between the elastic net models and the corresponding criteria were: $R_{Education} = .51$; $R_{Health} = .48$; $R_{BMI} = .43$; $R_{SleepQuality} = .42$; and $R_{SmokingFrequency} = .33$. On average across the five criteria, the random forest was

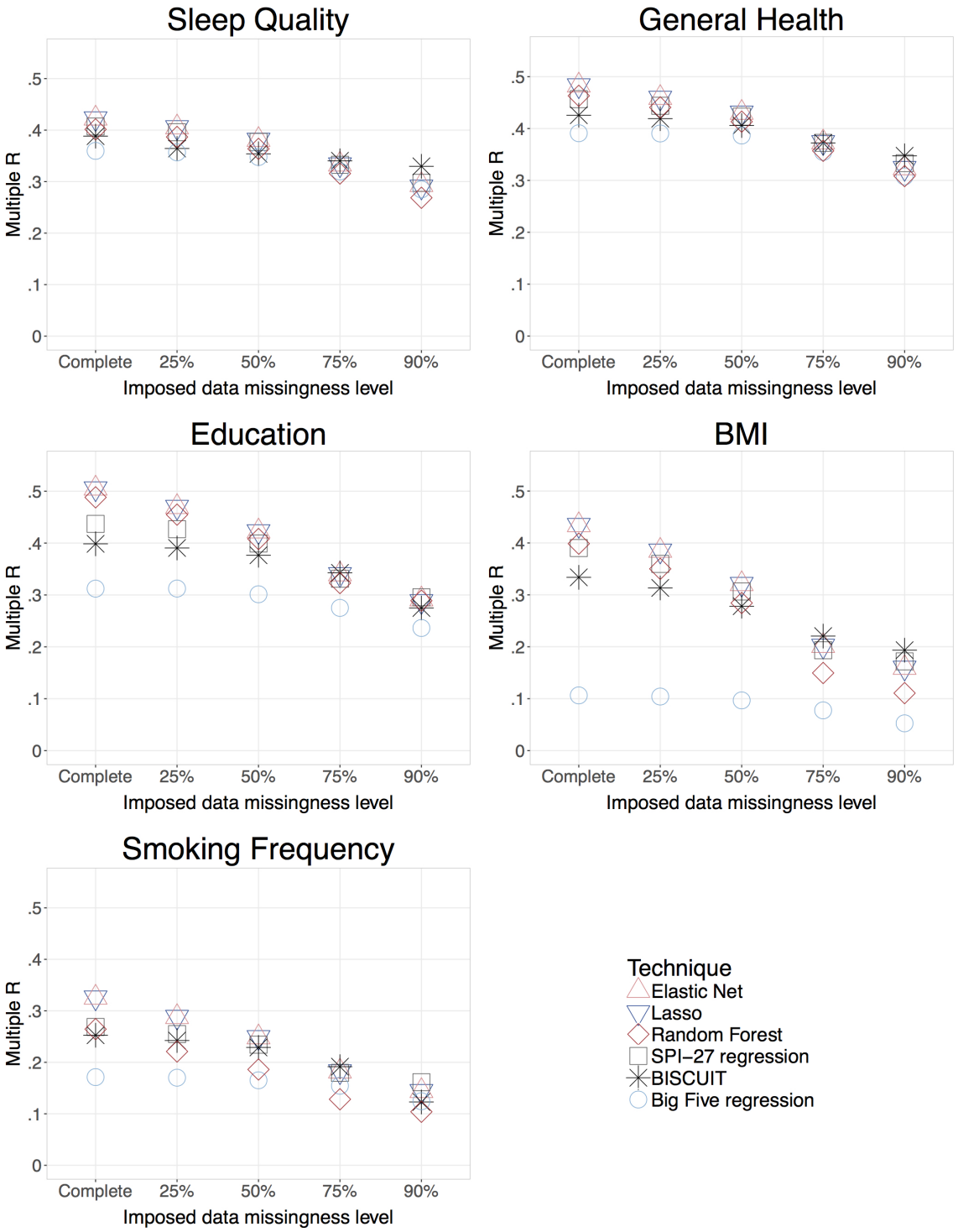
the 3rd most predictive technique for complete data, being 85% as predictive as the elastic net; regression using the SPI-27 (4th) was 81% as predictive; BISCUIT (5th) was 69% as predictive; and regression using the Big Five (last) was 42% as predictive.

One aim of the study was to determine whether BISCUIT, relative to other models, gained an advantage in predictive accuracy as data missingness increased. To assess this question, a ratio was found by dividing the accuracy of each BISCUIT model in each condition by the accuracy of the most predictive model in that condition, and these ratios were averaged for each level of data missingness. Consistent with our hypothesis, each increased level of missingness resulted in an improvement to BISCUIT's average comparative predictive accuracy, up to 75% data missingness: for complete data and 25%, 50%, and 75% data missingness, BISCUIT was, on average, 69%, 74%, 83%, and 100% as predictive as the most predictive model, respectively. In the 75% data missingness condition, BISCUIT had the highest predictive accuracy for four of the five criteria. In the 90% data missingness condition, BISCUIT's comparative predictive accuracy was, on average, 89% as predictive as the most predictive model, and BISCUIT had the highest predictive accuracy for three criteria.[5] The comparative predictive accuracy of regression using the SPI-27 also improved as data missingness increased: in the 90% data missingness condition, regression using the SPI-27 had the highest predictive accuracy for two criteria.[6]

[5]We also ran BISCUIT on imputed data to estimate a possible effect of noise generated by imputation. The predictive accuracy of BISCUIT using imputed data was 94% as predictive as BISCUIT using missing data, in terms of R^2 (see Table A.13 in Appendix A).

[6]A reviewer was concerned that the superiority of regression using the SPI-27, in the 90% data missingness condition and for the two criteria, was due to regression's tendency to capitalize on chance. They suggested that a model that aggregated regression coefficients across 10 folds would be more stable and less predictive, such that an aggregated regression model using the SPI-27 would not have the highest predictive accuracy for any of the criteria in the 90% data missingness condition. This hypothesis was tested and the results were null: across the five criteria in the 90% data missingness condition, the mean absolute difference in multiple R between the two regression methods was .0008, and the aggregated regression model using the SPI-27 was still the most predictive for the two criteria.

Figure 2.2: Predictive accuracy (measured in multiple R) of the six statistical techniques, using personality data, across five levels of imposed data missingness, in five criteria.



2.4.2 Parsimony

Parsimony of SL models was measured by the number of items used in a model; models that used fewer items were more parsimonious. BISCUIT generated the most parsimonious SL model in 23 of the 25 total conditions (Table A.3 in Appendix A).[7] The lasso generated the most parsimonious SL model in two of the 25 conditions (Table A.4 in Appendix A). SL techniques were ranked for their overall parsimony by calculating the mean and median number of items used in their models across the 25 conditions. Across the 25 conditions, BISCUIT was the most parsimonious technique, using, on average, 30 personality items per model ($median = 30, SD = 22, range = 1$–$81$); the lasso (2nd) used an average of 59 items per model ($median = 56, SD = 27, range = 14$–$112$); the elastic net (3rd) used an average of 60 items per model ($median = 58, SD = 27, range = 15$–$113$; Table A.5 in Appendix A); and the random forest (last) used 135 items in every model. The lasso and elastic net used fewer items as missingness increased, whereas the BISCUIT did not.

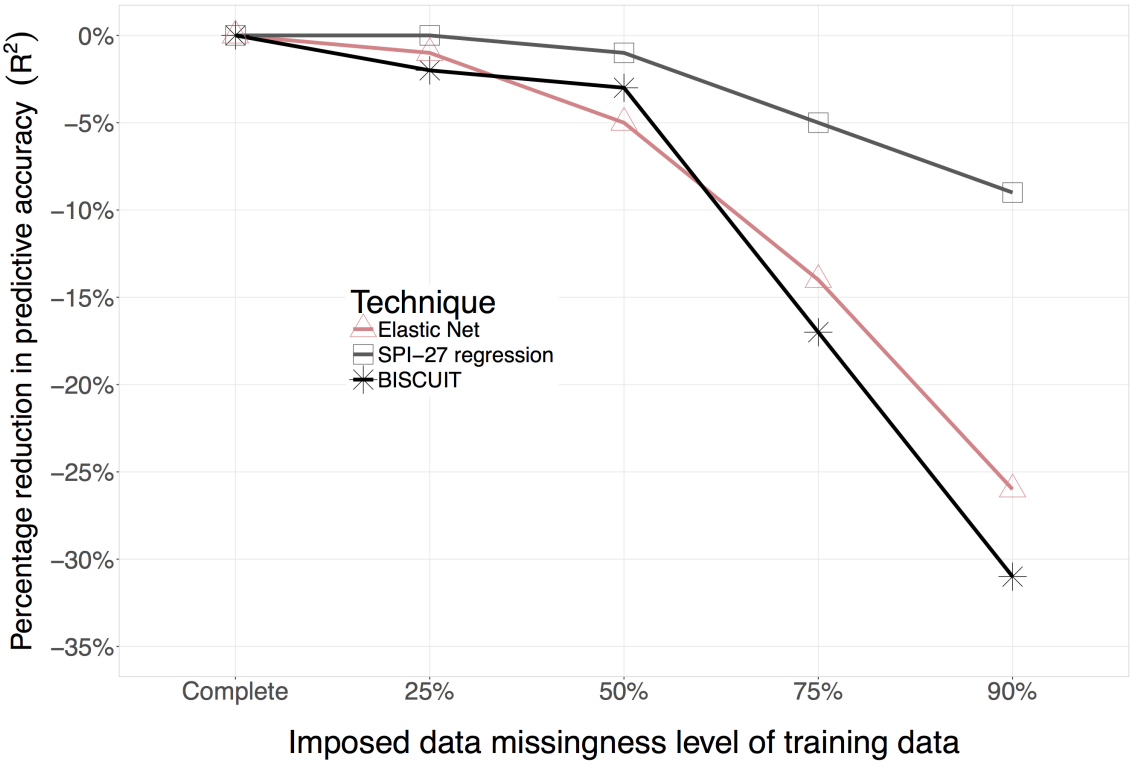
2.4.3 Post-hoc Analysis

Training models on data missingness conditions and testing them on complete data.

In the planned analyses, the predictive accuracy of each technique decreased as the amount of data missingness increased (Figure 2.2). This decrease in predictive accuracy was a combination of two effects: (a) the missingness in the training data, which gave each technique less information with which to build its predictive models; and (b) the missingness in the test data, which gave each technique less information with which to test its predictions. To isolate the first effect, we performed a post-hoc analysis to determine the decrease

[7] Of note is the fact that BISCUIT generated six one-item models in the 75% and 90% data missingness conditions. Five of these one-item models also had the highest predictive accuracy for their condition (Tables A.10 – A.12 in Appendix A). See Tables A.6 – A.8 in Appendix A for the item content of three brief BISCUIT models, each predicting a different criterion.

Figure 2.3: Percentage reduction in predictive accuracy (R^2) for each of three techniques, averaged across five criteria. Each model was trained on one of five levels of imposed data missingness and tested on complete data.



in predictive accuracy of models trained with data missingness but tested on complete data. We selected three techniques: the elastic net, regression using the SPI-27, and BISCUIT. Results indicated that the decrease in predictive accuracy due to missingness in training data was modest (Figure 2.3; Figure A.2 and Tables A.15 and A.16 in Appendix A). Loss in predictive accuracy was particularly low at the 50% data missingness condition; on average across the five criteria and three techniques, models trained on 50% data missingness were 95% as predictive as their respective models trained on complete data.

SL techniques on the SPI-27.

In the planned analyses, regression using the SPI-27 performed well across missingness levels and criteria. Because SL techniques are supposed to be an improvement over simple regression, we performed a post-hoc analysis to determine whether the predictive accuracy of models utilizing the SPI-27 could be improved with either of two SL techniques: the elastic net (the most predictive technique) and BISCUIT (the technique of special interest in this study). Results indicated that the predictive accuracy of models using the SPI-27 was not improved with the use of an SL technique instead of simple regression (Table A.14 in Appendix A).

2.5 Discussion

BISCUIT

Consistent with our hypothesis, the predictive accuracy of BISCUIT was more competitive with other SL techniques as data missingness increased, up to 75% data missingness, where it generated the model with the highest predictive accuracy in four of five criteria. BISCUIT did not perform as well in the 90% data missingness condition, but it generated the model with the highest predictive accuracy in three of the five criteria. Also consistent with our hypothesis, BISCUIT provided the most parsimonious model in 23 of 25 conditions.

The Elastic Net and Lasso

In terms of predictive accuracy, the elastic net dominated other techniques for the complete data and 25% and 50% data missingness conditions. The lasso was nearly as predictive as the elastic net. The elastic net and lasso may have dominated BISCUIT because BISCUIT's methodology ignored information that the elastic net and lasso did not. Specifically,

BISCUIT selected fewer variables than either technique, and BISCUIT used unit-weighting coefficients while the other two techniques used penalized regression coefficients.

The Random Forest

The random forest performed competitively for many missingness conditions and criteria. For complete data, it was 85% as predictive as the elastic net. It is possible that adjusting tuning parameters for the random forest could have increased its predictive accuracy, but we did not find a consensus in the literature regarding what, if any, tuning parameters should be used (Probst and Boulesteix, 2018; Tang et al., 2018). Increasing the number of trees per forest also may have helped, but the random forest was already the most burdensome SL technique in terms of computational load. The random forest appeared to be a lackluster choice for statistical learning with personality data, due to its suboptimal predictive accuracy, poor parsimony of its models, ambiguities in the literature regarding its tuning parameters, and its burdensome computational load.

Regression Using the SPI-27

Regression using the SPI-27 had greater predictive accuracy than the Big Five (for complete data, it was 93% more predictive), but in most conditions it did not have the maximal predictive accuracy of the elastic net. The SPI-27's dominance over the Big Five is consistent with previous research that found that narrower traits out-predicted broader traits (e.g., Paunonen and Ashton, 2001; Paunonen et al., 2003; Gladstone et al., 2019). In the 90% data missingness condition, regression using the SPI-27 had the most predictive model for two of five criteria. In such extreme data missingness, the benefit of improving the signal by aggregating items into facet-size factors may outweigh the benefit of utilizing item-level variance in a model's prediction. A post-hoc analysis indicated that the predictive accuracy of the SPI-27 was not improved by employing a more complex SL technique instead of simple

regression.

Regression Using the Big Five

As expected, regression using the Big Five had poor predictive accuracy compared to other techniques. For complete data, the Big Five was, on average, the least predictive technique of the six tested, being 42% as predictive as the elastic net. In no condition was regression using the Big Five the most predictive model. Additionally, regression using the Big Five showed a far weaker relationship between personality and BMI than any other technique (Figure 2.2; Table A.12 in Appendix A). This is consistent with previous findings in which analysis with broader traits failed to find personality-criterion relationships that were evident with narrower traits (Terracciano et al., 2009; Credé et al., 2017). If personality researchers continue to use the Big Five to answer the question, "Is personality related to this phenomenon," they may falsely conclude that no relationship exists, when narrower traits would have shown a robust relationship. Thus, regression or correlation using the Big Five may only be appropriate for studying personality-criterion relationships when no alternative is feasible.

Data Missingness

Across all techniques and criteria, predictive accuracy decreased as data missingness increased. However, a post-hoc analysis indicated that, after accounting for data missingness in the test data, loss in predictive accuracy was modest. That is, a model trained on a data set with missing or imputed data is still accurate, but complete data is needed to test this accuracy. Results indicated that the loss in predictive accuracy was approximately 5% for the 50% data missingness condition, which suggests that a large-sample study could introduce 50% data missingness without substantially impacting prediction. Fifty percent data missingness would allow for an item pool twice that of a complete data set, holding the

number of items per participant constant. Ninety percent data missingness would allow for an item pool ten times that of a complete data set, but the cost to predictive accuracy would be higher (this study estimated the range of loss to be approximately 10–30%). This loss in predictive accuracy will appear to be even greater if models are not tested on complete data. Thus, whether higher levels of data missingness are optimal for maximizing predictive accuracy will depend on whether the increased predictive accuracy due to a broader item pool will outweigh the loss due to data missingness.

2.5.1 Limitations of the Study

There were at least four methodological decisions that could impact the generalizability of the study's results. First, the comparative predictive accuracy of SL techniques may have depended upon the particular criteria or item pool; new criteria or item pools may favor different SL techniques. Second, only four SL techniques were compared in this study, and only one of them accounted for interactions (the random forest). Other SL techniques, such as Multivariate Adaptive Regression Splines (MARS; Friedman, 1991), may have better accounted for interactions than the random forest did. Third, the criteria chosen in this study were all assumed to be monotonic variables. Results related to the predictive accuracy of BISCUIT cannot be extended to non-monotonic criteria. Fourth, results for this study were based upon MMCAR data and may not generalize to data sets with non-random missingness, such as Missing Not At Random data sets.

Another major limitation of this study is that it compared the predictive accuracy of nuances with higher-order traits using an item pool in which all items were subsumed under higher-order traits. The scales of the SPI-27 (and scales which have followed classic psychometric internal consistency procedures) were designed such that the items were nothing more than representations of a scale; a personality scale does not include items that predict

outcomes well but are not exemplars of the scale. Thus, this study may have underestimated the predictive accuracy of nuance-based approaches, given a broader item pool.

2.5.2 Future Directions

Replication and Generalizability of Specific SL Models

Compared to traditional methods of analysis in personality psychology, statistical learning appears to be a more accurate approach to predicting criteria. The success of SL approaches is partially due to modeling the unique variance of personality items, which is ignored in higher-order traits. The superior predictive accuracy of SL techniques seems to suggest that domain-level personality-criterion relationships may be better described as a complex web of nuance-level patterns (e.g., Mõttus, 2016). But how stable are these patterns across data sets? In this study, an elastic net model best predicted BMI in the complete data condition, and this model contained 78 predictors and regression weights. Although the elastic net and other SL techniques did not capitalize on chance fluctuations and outliers, they may have capitalized on idiosyncratic attributes of this data set. A vital question to answer is: how predictive of a criterion is any specific SL model in a new data set that has different data collection methods, demographics, or other attributes? Another question to consider is: on average, how similar are two SL models generated from the same technique, using the same pool of predictors, but trained on substantially different data sets? Further research will be required to determine the generalizability of any given SL model, and whether parsimonious SL models are more replicable than complex SL models.

Utilizing a Planned Missing Data Structure to Train Statistical Learning Models

Post-hoc analysis indicated that there was relatively low cost to predictive accuracy for models trained on data sets with missingness, compared to models trained on complete

data. In the case of 50% data missingness, loss in predictive accuracy was about 5%. This finding suggests that researchers should consider using planned data missingness in their study designs. Randomly sampling items from a pool, instead of administering the same items to every participant, would allow a study to multiply the number of items in its pool while still allowing for the development of robust statistical learning models. In order for a model trained on MMCAR data to have maximal accuracy in predicting a criterion in a new data set, one would need to collect complete data on the variables that were included in the model. Of the techniques in this study, BISCUIT tended to have the fewest variables in its models, and in some models it had as few as one predictor (Table A.3 in Appendix A). Because it is an accurate, parsimonious and cost-effective statistical learning technique, BISCUIT could prove to be especially useful in applying personality-criterion models to real-world predictions of criteria.

2.5.3 Conclusions

Results from this study indicate that statistical learning techniques could prove to be essential in future research of personality-criterion relationships. SL techniques are low-cost tools that increase the predictive power of personality beyond traditional techniques; greater predictive accuracy is achieved by utilizing the same raw data. Since statistical learning methods excel at modeling item-level variance, item pools that contain a broad array of personality nuances may be more highly valued in the future. Planned data missingness designs are suited to meet the need for larger item pools; a study can collect data on an item pool of virtually any size, while still administering a given number of items per participant. Although both SL techniques and planned data missingness are powerful procedures, both can add complexity to a study. Statistical learning techniques such as BISCUIT offer a balanced approach to the study of personality-criterion relationships, by generating

parsimonious models that have greater predictive accuracy than traditional methods.

Chapter 3

Laying Personality BARE

Behavioral Frequencies Strengthen Personality-Criterion Relationships

3.1 Abstract

Personality consists of stable patterns of cognitions, emotions, and behaviors, yet behaviors are rarely studied in the field of personality psychology. Even when examined, behaviors typically are considered to be validation criteria for traditional personality items, instead of measures of personality. In the current study ($N = 332,489$), we conceptualize behavioral frequencies (self-reported yearlong patterns) as measures of personality. We investigate whether behavioral frequencies have incremental validity over and above traditional personality items in correlating personality with six outcome criteria. We use BISCUIT, a statistical learning technique, to find the optimal number of items for each criterion's model, across three pools of items: traditional personality items ($k = 696$), behavioral frequencies ($k = 425$), and a combined pool. Compared to models using only traditional personality items, models using

the behavioral frequency item pool are more strongly correlated to two criteria, and models using the combined pool are more strongly correlated to four criteria. We find mixed evidence that there is congruence between the type of criterion and the type of personality items that are most strongly correlated with it (e.g., behavioral criteria are most strongly correlated to behavioral personality items). Findings suggest that behavioral frequencies are measures of personality that provide a unique effect in describing personality-criterion relationships, over and above traditional personality items. We also provide an updated, public-domain item pool of behavioral frequencies: the BARE (Behavioral Acts, Revised and Expanded) Inventory.

3.2 Introduction

Of the three patterns commonly associated with personality (cognitions, emotions, and behaviors),[1] behaviors are the least studied (Baumeister et al., 2007; Furr, 2009). The measurement of cognitions and emotions have adequate coverage in traditional personality inventories, such as the NEO-PI-R (Costa and McCrae, 2008), the BFI-2 (Soto and John, 2017), the IPIP-HEXACO (Ashton et al., 2007), and the SPI-27 (Condon, 2018). Within each inventory, traditional personality items prompt participants to report how accurately trait descriptors (e.g., "I have a vivid imagination") describe a target, using a Likert-like scale. Some traditional personality items contain behavioral content; Wilt (2014) found that, of the Big Five domains, Conscientiousness and Extraversion had the most prototypically behavioral items (e.g., "Get chores done right away" and "Talk to a lot of different people at parties," respectively). Traditional personality items, however, (a) only include behaviors that are supposed to be indicative of broader personality traits; (b) do not measure specific frequencies of behaviors; and (c) do not specify a time period in which the behaviors have

[1] Wilt and Revelle (2015) have argued for a fourth pattern to be included: desires.

taken place.

The Act Frequency Approach (AFA; Buss and Craik, 1980, 1983, 1985, 1987) popularized the behavioral frequency, a retrospective, self-reported[2] number of instances that a target has performed a given behavior (e.g., meditated, littered) in a previous period of time (e.g., the past year). Compared to traditional personality items, behavioral frequencies may be more comprehensive and precise measurements of behavior because they (a) include a broader range of behaviors; (b) quantify the frequency of each behavior; and (c) specify a time period for each frequency of behavior. Despite the potential advantages that behavioral frequencies may have over traditional personality items, they have rarely been studied since the AFA was met with criticism in the late 1980s (e.g., Block, 1989; Moser, 1989). The babies (behavioral frequencies) were thrown out with the bath water (AFA's theory). Although there is a historical association between the AFA and behavioral frequencies, the use of behavioral frequencies does not require the baggage of AFA theory (namely, that there is no explanatory power in personality traits because they are merely behavioral summaries; for a review of AFA theory, see Buss and Craik, 1983).

Behavioral frequencies should be thought of as non-traditional personality items. As evidence for the claim that behavioral frequencies measure personality, we submit the following argument: (a) behavioral frequencies measure patterns of behavior; (b) personality includes patterns of behavior; (c) therefore, behavioral frequencies measure personality. Only statement *a* is debatable; *b* is widely accepted in personality psychology, and *c* follows directly from *a* and *b*. Block (1989) argued that behavioral frequencies do not measure behavior because "the indisputable fact remains that nowhere have acts been directly observed" (p. 237). Block's statement is accurate in the sense that behavioral frequencies typically are not observed and tallied by a third-party rater, but it ignores the fact that the behaviors have

[2]Behavioral frequencies may be reported by an informant, but previous studies have focused on self-reports. We use the term "behavioral frequency" as shorthand for self-reported behavioral frequency, unless otherwise noted.

been observed by a reporter who is intimately familiar with them: the participant.[3] Preliminary research suggests that retrospective self-reported behaviors are valid; self-reported behaviors and actual behaviors are positively related to one another (Gosling et al., 1998; Vazire and Mehl, 2008; Jackson et al., 2010).

Post-AFA researchers have conceptualized behavioral frequencies as validation criteria for traditional personality traits. For example: Grucza and Goldberg (2007) selected behavioral frequencies as one of several criteria to test the comparative validity of eleven personality inventories; Hirsh et al. (2009) found behavioral patterns for two metatraits (higher-order traits that supposedly subsume the Big Five); Church et al. (2007) found cross-cultural consistency of associations between behavioral frequencies and Big Five traits; Chapman and Goldberg (2017) described behavioral "signatures" of each Big Five trait; and Skimina et al. (2019) linked a person's values with their behavior. Since the AFA, however, no published paper has considered that behavioral frequencies are themselves measures of personality and that behavioral frequencies may account for variance in real-world criteria that is unexplained by traditional personality items.

3.2.1 Two Pilot Studies Have Examined the Incremental Validity of Behavioral Frequencies

Although personality psychologists typically study multi-item scales that represent broad traits, such as domains and facets, research suggests that individual items (sometimes called "nuances;" McCrae, 2015) also may be used to measure personality. Items are reliable measures; they are stable over time (Mõttus et al., 2017, 2019) and there is cross-rater agreement concerning their specific variance (Mõttus et al., 2014). Given the same pool of items, item-

[3] Additionally, it appears that Block was unaware of or ignored experience-sampling procedures (Csikszentmihalyi and Larson, 1987), in which individuals frequently report their current behaviors throughout the day.

based models better predict outcomes than models of multi-item scales, because the variance associated with individual items has predictive validity (Seeboth and Mõttus, 2018; Mõttus et al., 2015, 2020; Revelle et al., 2020).

Two unpublished pilot studies have compared the predictive validity of yearlong behavioral frequencies over traditional personality items. The first study ($N = 31,467$; Elleman et al., 2017), using 199 behavioral frequencies and 100 traditional personality items, found the ten personality items with the largest absolute correlation for each of four life outcomes. The items that most strongly correlated with criteria were overwhelmingly behavioral frequencies; of the total top 40 items (10 items multiplied by four criteria), only one was a traditional personality item. The second pilot study (Elleman et al., 2018) was a replication and extension of the first. It included more participants ($N = 177,853$), criteria (twelve), and personality items (696 traditional and 454 behavioral). Of the top 120 items (i.e., the 10 items with the largest absolute correlation for each of twelve criteria), 79 of them were traditional personality items. Overall, behavioral frequencies did not better predict criteria than traditional personality items, but they were represented in proportion to the size of their item pool.

Post-hoc analysis of the second study uncovered a pattern: in general, each criterion was predicted by mostly one type of personality item. Three criteria (body mass index [BMI], smoking frequency, and caffeine consumption) were predicted by behavioral frequencies, while four criteria (overall stress, general health, sleep quality, and prescription adherence) were predicted by traditional personality items. One criterion, exercise frequency, was predicted by an even mix of the two personality item types. Models were not able to produce acceptable cross-validated predictions for the remaining four criteria (frequency of brushing and flossing teeth, hospital emergency room (ER) visits, and average hours slept). Qualitative analysis indicated that three of the four criteria that were predicted by traditional personality items appeared to be more similar to trait descriptors than measures of behavior: health

("How would you rate your health?" Poor – Excellent); sleep quality ("How is the quality of your sleep?" Poor – Excellent); and stress ("How would you rate your stress lately?" Extremely calm – Extremely stressed). Prescription adherence ("Do you take medication as prescribed?" I often miss a dose – I never miss a dose) was a measure of behavior, but not especially precise. Interestingly, many of the best traditional personality items that predicted prescription adherence appeared to be a mix of behavior and cognition or emotion (e.g., "Quickly lose interest in the tasks I start"; "Do things that I later regret"; "Habitually blow my chances"; and "Do things without thinking of the consequences").

Conversely, the three criteria predicted by behavioral frequencies were precise measures of behavior or outcomes driven mostly by behavior: smoking frequency ("How often do you smoke?" Never in my life – More than 20 times a day); caffeine consumption ("How much caffeine do you consume each day?" None – More than 400mg a day); and BMI (a ratio of weight and height). These findings suggest that behavioral frequencies may better predict behavioral criteria that are precisely measured, while traditional personality items (i.e., cognitions and emotions) may better predict criteria that are more cognitive and emotional. Simply put, behaviors predict behaviors, while cognitions and emotions predict cognitions and emotions.

3.2.2 Overview of the Current Study

The primary aim of the current study was to determine, for six criteria, the extent to which behavioral frequencies accounted for additional variance above traditional personality items. To take a more empirical approach than the pilot studies, which selected an arbitrary number of items for a model, this study used a new statistical learning technique, BISCUIT (Revelle, 2020; Elleman et al., in press; see also Chapter 2), to determine the optimal number of personality items for each criterion. A secondary aim was to determine if the *post-hoc*

findings from the second pilot study could be replicated. That is, were some criteria mostly predicted by one type of personality item, and if so, was this type of personality item congruent with the type of criterion (i.e., were behavioral criteria predominantly correlated with behavioral frequencies, and were cognitive and emotional criteria predominantly correlated with traditional personality items)? Lastly, in the current study we released an updated, public-domain item pool of behavioral frequencies: the BARE (Behavioral Acts, Revised and Expanded) Inventory (Table B.1 in Appendix B).

3.3 Methods

3.3.1 Participants

Participant data were collected from https://SAPA-project.org as part of the Synthetic Aperture Personality Assessment (SAPA) project (Revelle et al., 2016). Participants received automated feedback regarding their personality as compensation for their participation. The data collection time period for this study (May 2018 to November 2019) started immediately after the second pilot study. Participants were included in the study if they responded to at least one behavioral frequency item. Participants ($N = 332,489$) were from 230 countries, 65% were female, and the median age was 29 years ($min = 14$, $max = 90$, *median absolute deviation* $= 15$). Of the 85% of participants who reported their educational attainment, 18% were enrolled in college and 56% had attained at least an associate's degree. Participants from the United States accounted for 51% of the sample. Of the 61% of U.S. participants who reported their ethnicity, 78% identified as White, 7% as Hispanic American, 4% as African American, 4% as Asian American, 1% as Native Alaskan/Hawaiian/American, and 6% as multi-racial, "other," or "none of these."

3.3.2 Measures

MMCAR Structure

Because there were thousands of personality items in SAPA's item pool, each participant received a quasi-random sample of them; each inventory may have been sampled at a different rate, but within each inventory, a random sample of items was given. This data collection method resulted in a Massively Missing Completely at Random (MMCAR) data structure where, for any given participant, most of the data were missing (Revelle et al., 2010, 2016). This MMCAR approach was not used for demographic or criterion variables; every participant was given all of those items, although participants were not required to respond to them.

Traditional Personality Items

There were 696 traditional personality items included in this study. These items are public domain and have been curated for use on the SAPA website (Condon and Revelle, 2015; Condon et al., 2017). The items were from eight sets of personality scales, seven of which were from the International Personality Item Pool (IPIP; `http://ipip.ori.org/`), a repository of public domain items (Goldberg, 1999; Goldberg et al., 2006). Items from the following inventories are mentioned in the Results section: IPIP-NEO (Goldberg, 1999); IPIP-HEXACO (Ashton et al., 2007); QB6 (Thalmayer et al., 2011); BFAS (DeYoung et al., 2007); EPQ (Eysenck et al., 1985); and Plasticity/Stability (DeYoung, 2010). All traditional personality items were given the same six-point Likert-like scale: "Very Inaccurate," "Moderately Inaccurate," "Slightly Inaccurate," "Slightly Accurate," "Moderately Accurate," and "Very Accurate." Compared to other traditional personality inventories, the 135 items of the SPI-27 (Condon, 2018), which are in some of the reported scales, were oversampled; the median number of administrations of an item from the SPI-27 was 225,563, whereas the

median number of administrations of a non-SPI traditional personality item was 2,878.

Behavioral Frequencies

There were 425 behavioral frequency items included in this study. These items constituted the BARE Inventory, which combined items from the Oregon Avocational Interest Scales (ORAIS; Goldberg, 2010) and items curated and revised from the Behavioral Acts Inventory (BAI; Chapman and Goldberg, 2017).[4] For each item, participants rated the frequency of their behavior on a six-point scale: "Never in my life," "Not in the past year," "Less than 3 times in past year," "3 to 10 times in past year," "10 to 20 times in past year," and "More than 20 times in past year." The median number of administrations of a behavioral frequency was 7,718.

Demographic and Criterion Variables

There were four self-reported demographics: age, sex, ethnicity, and educational attainment. The median number of administrations of a demographic variable was 268,452 (ethnicity had far fewer administrations [$k = 94,065$] due to only being applicable for participants in the United States). There were six self-reported criterion variables: general health ("How would you rate your health?" Poor – Excellent); overall stress ("How would you rate your stress lately?" Extremely calm – Extremely stressed); body mass index (computed from weight and height); exercise frequency ("How often do you exercise?" Very rarely or never – More than five times a week); smoking frequency ("How often do you smoke?" Never in my life – More than 20 times a day); and hospital emergency room visits ("How many times have you been admitted to an emergency room in the last 6 months?" None – Three or more times). The median number of administrations of a criterion was 175,138. Fewer criteria

[4]See Appendix B for a description of how the BARE Inventory was created. See Table B.1 in Appendix B for a list of items in the BARE Inventory.

were included in this study than the second pilot study due to a data sharing agreement with the administrator of the SAPA website.

3.3.3 Statistical Analyses

All analyses were performed in the statistical programming language R (R Core Team, 2019), using the RStudio environment (RStudio Team, 2019). Due to the MMCAR data structure, it was not appropriate to use the full sample size ($N = 332,489$) to estimate statistical significance. We took a conservative approach for determining the effective n for each analysis related to a criterion by finding the minimum number of pairwise administrations of a pool of items with a criterion. For example, the minimum number of pairwise administrations between the 696 traditional personality items and the general health criterion was 2,410, so this number was used as the effective n for determining the statistical significance of analyses for general health. Separately, for a general estimate of statistical significance for item-level correlations, we calculated the minimum absolute correlation that would be statistically significant ($p < .05$), using the fewest number of pairwise administrations with a criterion ($n = 1,736$) and a Bonferroni correction (Dunn, 1961) for the maximum number of correlations between a criterion and personality items ($k = 1,121$). All item-level personality-criterion correlations across all BISCUIT models were greater than or equal to this threshold ($|r| = .10$).

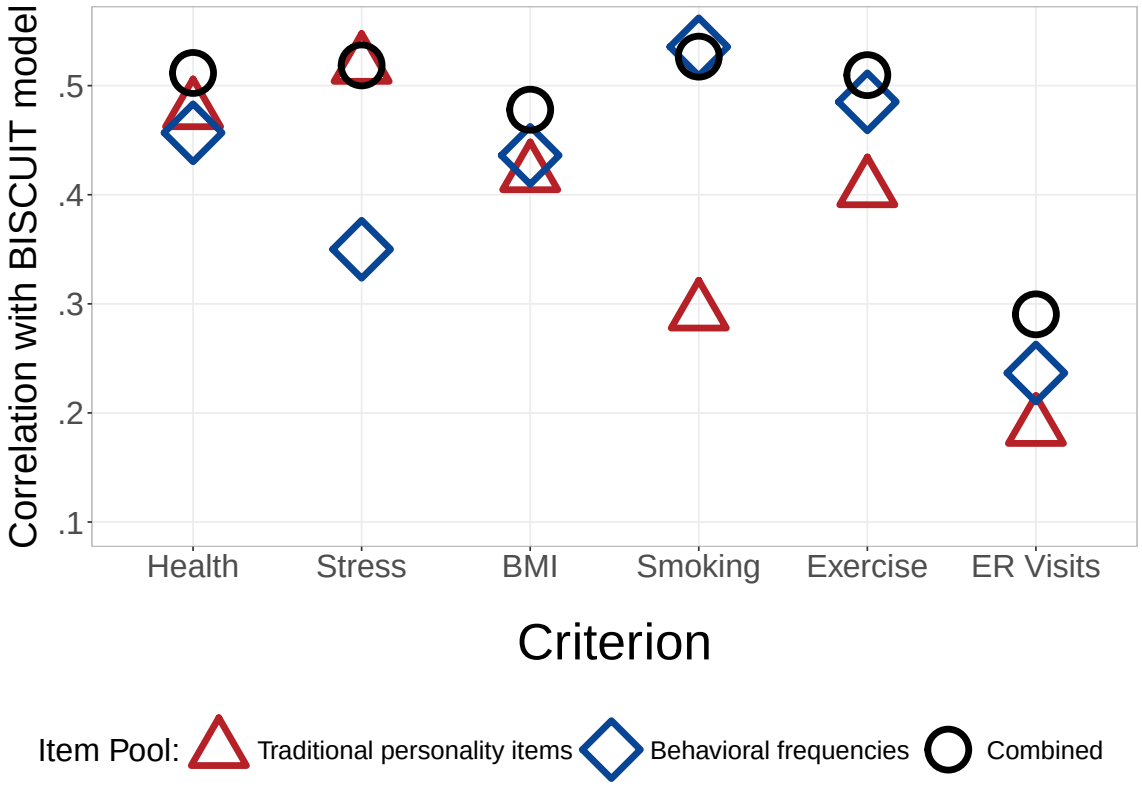
The statistical learning technique BISCUIT (Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent) was used to find a list of items that were most strongly correlated with each criterion. BISCUIT used a k-fold resampling procedure ($k = 10$) to determine a cross-validated list of the "best items" based upon each item's average correlation with the criterion (for a concise description of k-fold, see Chapman et al., 2016, p. 60). One unique feature of BISCUIT is that its models implement unit-weighting to reduce

overfitting. We employed BISCUIT in this study because it was designed for MMCAR data structures; BISCUIT calculates item-level correlations from the pairwise administrations of items and does not need to impute missing data, unlike some other SLTs. The BISCUIT algorithm is available as the "bestScales" function in the "psych" R package (Revelle, 2020, version 2.0.5).

We gave BISCUIT a limited range of possible solutions for the optimal number of items in a model: from 10–100 items. The minimum of 10 items was chosen in order to: (a) have a large enough frequency of items in each model for chi-squared tests by item type and (b) ensure a reasonable number of items per model for assessing item content. The maximum of 100 items was chosen in order to: (a) decrease processing time for computing results, (b) select an arbitrary number large enough to be considered outside the realm of a parsimonious "best items" solution. We gave BISCUIT a preference for more parsimonious models by having it select the model with the fewest number of items that was within one standard error of the optimal model, since these two models would be statistically no different from one another in terms of their correlation with a criterion. For each criterion, BISCUIT found the best personality items using three pools of items: (a) traditional personality items, (b) behavioral frequencies, and (c) a combined set of all personality items.

After a BISCUIT model was built and cross-validated on a criterion, we examined the item content of the best items. Any of the best items that were synonymous with the criterion were removed from the pool of possible items and the model was rerun. For example, the behavioral frequency "smoked tobacco" was removed from the item pool for the criterion "smoking frequency." Across the six criteria, twelve personality items were removed. For a complete list of removed personality items, see Table B.2 in Appendix B. For the reliabilities of each BISCUIT model as if it were a typical personality scale, see Table B.3 in Appendix B.

Figure 3.1: Correlation of BISCUIT models with six criteria, using three pools of personality items (traditional items, behavioral frequencies, and a combined pool). The height of each shape is approximately the size of the estimate's 95% confidence interval.



3.4 Results

3.4.1 The Strength of Personality-Criterion Relationships Using Different Item Pools

Did behavioral frequencies strengthen personality-criterion relationships beyond that of traditional personality items? To answer this question, we tested for differences in non-independent correlations (Steiger, 1980) of a criterion and BISCUIT models using different item pools. For this analysis, p-values were Holm-adjusted (Holm, 1979) to account for the

total of 12 comparisons. First, we determined whether there were any instances in which BISCUIT models built with behavioral frequencies were more strongly correlated with criteria than models built with traditional personality items. Two criteria (smoking and exercise frequency) were more strongly correlated with BISCUIT models built with behavioral frequencies than those built with traditional personality items. One variable, overall stress, was more strongly correlated with the BISCUIT model built with traditional personality items. And each of the remaining three criteria (general health, BMI, and ER visits) was not differentially correlated with BISCUIT models built with the two item types (Figure 3.1; Table 3.1). Second, we determined whether there were any instances in which BISCUIT models built with all personality items were more strongly correlated with criteria than models built with traditional personality items. For these comparisons, we adjusted the correlations between each pair of BISCUIT models to account for the fact that they had overlapping items,[5] which had the effect of more conservative estimates of statistical significance. Four out of the six criteria were more strongly correlated with BISCUIT models built with the combined item pool than those built with traditional personality items; BISCUIT models for overall stress and BMI were not improved by using the combined item pool (Figure 3.1; Table 3.1).

3.4.2 The Types of Personality Items Most Related to Each Criterion

For each criterion, were the personality items selected by BISCUIT predominantly of one type? To answer this question, for each criterion's best items, we used Pearson's chi-squared tests to determine if frequencies of item types were statistically different than the expected distribution, in which 696 (62%) were traditional personality items and 425 (38%) were behavioral frequencies. Due to the small number of each criterion's best items, statistical

[5]We used the "scoreOverlap" function of the "psych" package in R, which implements an algorithm similar to suggestions by Cureton (1966) and Bashaw and Anderson (1967).

Table 3.1: Tests to determine the differences in non-independent correlations of criteria with BISCUIT models that use traditional personality item pools, compared to other item pools. Each test determines if correlation r_{AB} is significantly different from r_{AC}, accounting for r_{BC}. Variables A are six criteria. Variables B are BISCUIT models built using the traditional personality item pool. Variables C are BISCUIT models built using either the behavioral frequency item pool or an item pool that combined both personality item types. Bolded p-values indicate significant differences ($p < .05$), and bolded item pools and correlations indicate the models with the larger correlations.

A: Criterion	B: BISCUIT item pool	Items used(B)	C: BISCUIT item pool	Items used(C)	r_{AB}	r_{AC}	r_{BC}	t value	p value
General health	Traditional	10	Behaviors	10	.48	.46	.44	1.11	.807
General health	Traditional	10	**Combined**	10	.48	**.51**	.87	−3.89	**<.001**
Overall stress	**Traditional**	20	Behaviors	10	**.52**	.35	.56	8.83	**<.001**
Overall stress	Traditional	20	Combined	20	.52	.52	.94	0.00	1.000
Body mass index	Traditional	41	Behaviors	10	.42	.44	.38	−0.72	.938
Body mass index	Traditional	41	Combined	14	.42	.48	.35	−2.51	.060
Smoking frequency	Traditional	26	**Behaviors**	10	.29	**.54**	.41	−11.11	**<.001**
Smoking frequency	Traditional	26	**Combined**	10	.29	**.53**	.45	−11.04	**<.001**
Exercise frequency	Traditional	27	**Behaviors**	10	.41	**.49**	.47	−3.78	**<.001**
Exercise frequency	Traditional	27	**Combined**	10	.41	**.51**	.80	−8.02	**<.001**
ER visits	Traditional	10	Behaviors	10	.19	.24	.25	−1.80	.290
ER visits	Traditional	10	**Combined**	10	.19	**.29**	.56	−4.93	**<.001**

significance was determined by a Monte Carlo simulation with 2,000 replicates (Hope, 1968). Results indicated that three of the six criteria were predominantly associated with one type of personality item, and each of those criteria was associated with the type of personality item that we expected: overall stress was predominantly associated with traditional personality items; and BMI and smoking frequency with behavioral frequencies (Table 3.2).

3.4.3 Summary of Best Items Content for Each Criterion[6]

Of the ten personality items selected by BISCUIT for correlating with general health, there were nine traditional personality items from four inventories (Table 3.3). Three traditional personality items were from the Liveliness facet of the Extraversion domain (e.g.,

[6]See Tables B.4 – B.15 in Appendix B for the best items content of BISCUIT models built only with traditional personality items or behavioral frequencies.

Table 3.2: Pearson's chi-squared tests comparing the frequencies of behavioral and traditional items in the total item pool against the frequencies in each BISCUIT model that was built with the total item pool. A bolded row indicates statistical significance ($p < .05$).

Criterion	Behavioral items	Traditional items	χ^2	p value
General health	1	9	3.29	.119
Overall stress	**0**	**20**	**12.08**	**.001**
Body mass index	**11**	**3**	**9.66**	**.003**
Exercise frequency	3	7	0.26	.748
Smoking frequency	**9**	**1**	**11.37**	**.001**
ER visits	7	3	4.32	.056
Total item pool	425	696	–	–

"Have great stamina"), three were from the Resiliency domain (e.g., "Recover quickly from stress and illness"), and three were from Neuroticism domain of two different inventories (e.g., "Have a low opinion of myself"). The one behavioral frequency was, "Did aerobic exercise."

Of the twenty personality items selected by BISCUIT for correlating with overall stress, there were twenty traditional personality items from six inventories (Table 3.4). Fifteen items were from the Neuroticism and Emotional Stability domains of four different inventories (e.g., "Feel desperate"), three were from the Resiliency domain (e.g., "Feel a sense of worthlessness or hopelessness"), one was from the Stability metatrait (i.e., "Find life difficult"), and one was from the Liveliness facet of the Extraversion domain (i.e., "Feel healthy and vibrant most of the time").

Of the fourteen items selected by BISCUIT for correlating with BMI, there were three traditional personality items from the Immoderation facet of the Neuroticism domain (e.g., "Often eat too much"; Table 3.5). All three of these items mentioned behaviors in relation to self-control (e.g. "Am able to control my cravings"). The other ten items were behavioral frequencies involving food (e.g., "Ate too much"), monitoring one's health (e.g., "Had my

Table 3.3: The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with general health ($R = .51$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet*	Key
Tire out quickly.	−.39	Traditional	IPIP-HEXACO	Ext./Liveliness	−
Have great stamina.	.38	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Am usually active and full of energy.	.36	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Often feel listless and tired for no reason.	−.34	Traditional	EPQ	Neuroticism	+
Recover quickly from stress and illness.	.34	Traditional	QB6	Resiliency	+
Am happy with my life.	.34	Traditional	QB6	Resiliency	+
Feel a sense of worthlessness or hopelessness.	−.32	Traditional	QB6	Resiliency	−
Have a low opinion of myself.	−.32	Traditional	IPIP-NEO	Neur./Depression	+
Feel that I'm unable to deal with things.	−.32	Traditional	IPIP-NEO	Neur./Vulnerability	+
Did aerobic exercise.	.31	Behavioral	BARE (ORAIS)	Exercise	+

*Ext. = Extraversion; Neur. = Neuroticism

Table 3.4: The 20 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with overall stress ($R = .52$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet*	Key
Find life difficult.	.41	Traditional	Plasticity/Stability	Stability	−
Am relaxed most of the time.	−.40	Traditional	IPIP-NEO	Neur./Anxiety	−
Feel a sense of worthlessness or hopelessness.	.37	Traditional	QB6	Resiliency	−
Feel desperate.	.37	Traditional	IPIP-NEO	Neur./Depression	+
Recover quickly from stress and illness.	−.37	Traditional	QB6	Resiliency	+
Am happy with my life.	−.37	Traditional	QB6	Resiliency	+
Am often down in the dumps.	.36	Traditional	IPIP-NEO	Neur./Depression	+
Often feel blue.	.36	Traditional	IPIP-NEO	Neur./Depression	+
Feel healthy and vibrant most of the time.	−.36	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Get caught up in my problems.	.36	Traditional	IPIP-NEO	Neur./Anxiety	+
Worry about things.	.36	Traditional	IPIP-NEO	Neur./Anxiety	+
Often feel fed-up.	.35	Traditional	EPQ	Neuroticism	+
Rarely feel depressed.	−.35	Traditional	BFAS	Neur./Withdrawal	−
Am often in a bad mood.	.35	Traditional	IPIP-NEO	Neur./Anger	+
Dislike myself.	.35	Traditional	IPIP-NEO	Neur./Depression	+
Often feel lonely.	.35	Traditional	EPQ	Neuroticism	+
Rarely worry.	−.35	Traditional	IPIP-HEXACO	EmS./Anxiety	−
Feel that I'm unable to deal with things.	.34	Traditional	IPIP-NEO	Neur./Vulnerability	+
Suffer from nerves.	.34	Traditional	EPQ	Neuroticism	+
Am a worrier.	.33	Traditional	EPQ	Neuroticism	+

*EmS = Emotional Stability; Ext. = Extraversion; Neur. = Neuroticism

cholesterol level checked"), or commuting and motor vehicles (e.g., "Used public transportation").

Table 3.5: The 14 personality items most strongly correlated with **body mass index**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with BMI ($R = .48$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet*	Key
Often eat too much.	.34	Traditional	IPIP-NEO	Neur./Immoderation	+
Ate too much.	.26	Behavioral	BARE (ORAIS)	Food-Related	+
Dieted to lose weight.	.26	Behavioral	BARE	None	
Had my cholesterol level checked.	.24	Behavioral	BARE	None	
Used public transportation.	−.24	Behavioral	BARE (ORAIS)	Green Activities	+
Consulted a professional nutritionist, dietician, or physician about my diet.	.24	Behavioral	BARE	None	
Am able to control my cravings.	−.24	Traditional	IPIP-NEO	Neur./Immoderation	−
Ate or drank while driving.	.23	Behavioral	BARE (ORAIS)	Food-Related	+
Took antacids.	.23	Behavioral	BARE	None	
Took three or more different medications in the same day.	.22	Behavioral	BARE	None	
Had my blood pressure taken.	.21	Behavioral	BARE	None	
Bought a car, truck, or motorcycle.	.21	Behavioral	BARE (ORAIS)	Vehicles	+
Rarely overindulge.	−.20	Traditional	IPIP-NEO	Neur./Immoderation	−
Drove a car 10 miles (16 km) per hour over the speed limit.	.20	Behavioral	BARE	None	

*Neur. = Neuroticism

Of the ten personality items selected by BISCUIT for correlating with smoking frequency, there was one traditional personality item from the Psychoticism domain (i.e., "Would take drugs which may have strange or dangerous effects"; Table 3.6). The other nine items were behavioral frequencies involving the use of drugs (e.g., "Smoked, vaped or otherwise consumed marijuana") or alcohol (e.g., "Became intoxicated").

Of the ten personality items selected by BISCUIT for correlating with exercise frequency, there were seven traditional personality items from four inventories (Table 3.7). Four traditional personality items were from the Liveliness facet of the Extraversion domain (e.g., "Am usually active and full of energy"), two were from the Neuroticism domain of two different

Table 3.6: The 10 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with smoking frequency ($R = .53$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet	Key
Smoked, vaped or otherwise consumed marijuana.	.50	Behavioral	BARE	None	
Drank alcohol or used other drugs to make myself feel better.	.37	Behavioral	BARE	None	
Took a hard drug recreationally (such as cocaine, methamphetamine, or heroin).	.33	Behavioral	BARE	None	
Would take drugs which may have strange or dangerous effects.	.32	Traditional	EPQ	Psychoticism	+
Had a hangover.	.32	Behavioral	BARE (ORAIS)	Drinking	+
Left a place because of cigarette smoke.	−.32	Behavioral	BARE	None	
Became intoxicated.	.28	Behavioral	BARE (ORAIS)	Drinking	+
Tried to stop using alcohol or other drugs.	.27	Behavioral	BARE	None	
Used smokeless tobacco (such as chewing tobacco or snuff).	.27	Behavioral	BARE	None	
Had an alcoholic drink before breakfast or instead of breakfast.	.25	Behavioral	BARE	None	

inventories (e.g., "Often feel listless and tired for no reason"), and one was from the Activity Level facet of the Extraversion domain (i.e., "Do a lot in my spare time"). The three behavioral frequencies involved behaviors that were related to an active lifestyle (e.g., "Went on a hike").

Of the ten personality items selected by BISCUIT for correlating with emergency room visits, there were three traditional personality items from two inventories (Table 3.8). Two items were from the Plasticity metatrait (e.g., "Find myself in the same kinds of trouble, time after time"), and one was from the Immoderation facet of the Neuroticism domain (i.e., "Don't know why I do some of the things I do"). The other seven items were behavioral frequencies, six of which involved health and medical behaviors (e.g., "Took three or more different medications in the same day"). The other behavioral frequency was, "Cried nearly every day for a week."

Table 3.7: The 10 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .51$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet*	Key
Went on a hike.	.37	Behavioral	BARE (ORAIS)	Summer Activities	+
Feel healthy and vibrant most of the time.	.35	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Am usually active and full of energy.	.33	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Tire out quickly.	−.32	Traditional	IPIP-HEXACO	Ext./Liveliness	−
Have great stamina.	.31	Traditional	IPIP-HEXACO	Ext./Liveliness	+
Often feel listless and tired for no reason.	−.30	Traditional	EPQ	Neuroticism	+
Took a long walk alone.	.30	Behavioral	BARE	None	
Do a lot in my spare time.	.29	Traditional	IPIP-NEO	Ext./Activity level	+
Am easily discouraged.	−.27	Traditional	BFAS	Neur./Withdrawal	+
Attended an athletic event.	.26	Behavioral	BARE (ORAIS)	Sports	+

*Ext. = Extraversion; Neur. = Neuroticism

Table 3.8: The 10 personality items most strongly correlated with **emergency room visits**, selected by BISCUIT from a pool of 1,121 items. The BISCUIT model composed of these items had a moderate correlation with emergency room visits ($R = .29$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$. The column "Key" indicates whether the item was positively or negatively keyed on the listed domain/facet.

Item	Corr.	Item pool	Inventory	Domain/Facet	Key
Had my blood pressure taken.	.17	Behavioral	BARE	None	
Took three or more different medications in the same day.	.16	Behavioral	BARE	None	
Visited a doctor for a physical examination or general check up.	.16	Behavioral	BARE	None	
Cried nearly every day for a week.	.13	Behavioral	BARE	None	
Don't know why I do some of the things I do.	.13	Traditional	IPIP-NEO	Neur./Immoderation	+
Find myself in the same kinds of trouble, time after time.	.13	Traditional	Plasticity/Stability	Stability	−
Changed my daily routine because of pain associated with an injury or illness.	.13	Behavioral	BARE	None	
Had a medical operation.	.12	Behavioral	BARE	None	
Used a thermometer to take my temperature.	.12	Behavioral	BARE	None	
Am self-destructive.	.12	Traditional	Plasticity/Stability	Stability	−

3.5 Discussion

Behavioral frequencies strengthened personality-criterion relationships beyond what was possible with traditional personality items.

Four of the six criteria in this study were more strongly correlated with a BISCUIT model built with a combined item pool than a BISCUIT model built with traditional personality items. This result indicates that behavioral frequencies have incremental validity for relating personality with criteria of interest; behavioral frequencies capture unique variance in the personality patterns of individuals. Additionally, two of these four criteria were more strongly correlated with a model built with behavioral frequencies than traditional personality items. In some cases, behavioral frequencies may be better than traditional personality items at establishing the strongest relationship between life outcomes and personality. Lastly, there was only one instance in which a BISCUIT model built with traditional personality items outperformed a model built with behavioral frequencies. Thus, in many cases researchers may be able to entirely replace traditional personality items with behavioral frequencies and have no detrimental impact to the predictive accuracy of personality-criterion models.

There was mixed evidence for congruence between the type of personality item and type of criterion.

In three of six BISCUIT models which used the combined personality item pool, criteria were related to predominantly one type of personality item; BMI and smoking frequency were predominantly correlated with behavioral frequencies, and overall stress was predominantly correlated with traditional personality items. For BMI and smoking frequency, the few traditional personality items in each model were behaviors that were consistent with the behavioral frequencies in that model. For instance, the traditional personality item most related to BMI was “Often eat too much,” which was synonymous with the behavioral

frequency "Ate too much." And the one traditional personality item related to smoking frequency, "Would take drugs which may have strange or dangerous effects," presented a hypothetical behavior which agreed with the actual behaviors in the model (e.g., "Took a hard drug recreationally, such as cocaine, methamphetamine, or heroin"). These results were aligned with our hypothesis that behaviors would predict behavioral criteria and that cognitions and emotions would predict cognitive and emotional criteria.

However, three criteria (general health, exercise frequency, and ER visits) were not predominantly associated with one type of personality item. This evidence suggests that some criteria may be mostly related to a personality item type that is congruent with its type, while some criteria may show no such pattern. Interestingly, there were no examples in this study of a criterion that was predominantly correlated with a personality item type that was incongruent with the criterion's type; no behavioral criterion was associated with mostly cognitive/emotional personality items, or vice versa. Of course, the absence of evidence in this study does not preclude the possibility that this type of personality-criterion incongruence exists. A full reckoning of our hypothesis concerning personality-criterion type congruence would require many more criteria, with perhaps an even larger personality item pool, as well as consideration for a more nuanced typology of criteria and personality items (e.g., separating cognitions and emotions into their own types, and/or incorporating desires as a type). Results from the current study suggest that researchers may want to more carefully consider the type of personality items that they select when associating criteria with personality.

Transparent SLTs can elucidate how personality is related to outcomes of interest.

A common criticism of statistical learning techniques is that they produce models that are uninterpretable "black boxes" (Yarkoni and Westfall, 2017). Some SLTs, like BISCUIT, however, output models that are transparent enough to let researchers peek inside. In the case of the criterion "emergency room visits," many of the behavioral items selected by the

BISCUIT model were behaviors which, on their face, appear to be the actions of someone in poor health who would be more likely to require a visit to the emergency room (e.g., "Took three or more different medications in the same day," and "Changed my daily routine because of pain associated with an injury or illness"). These particular behaviors seem to constitute a coherent pattern, but probably would not be useful for the development of a personality trait. One might even argue that these behaviors ultimately would not prove to be of much practical value to researchers; common sense could tell you that a person who has reported, "I had a medical operation" is also more likely to report having gone to an emergency room, perhaps even having had the medical operation as a result of going to the ER.

It is important to remember, however, that the purpose of this study was to highlight the predictive potential of behavioral frequencies, not construct a latent trait of emergency-room-ness. If the behaviors selected by an SLT are too similar to a criterion to be of practical use, researchers can prune those items from a model and perhaps also add a broader range of behaviors to the item pool. Some outcomes, such as being struck by a motor vehicle while enjoying dinner at a restaurant, may be so dependent upon environmental conditions that patterns of thoughts, feelings, and behaviors will not account for much of a criterion's variance, no matter the size of the item pool. However, even in the case of emergency room visits, there were three traditional personality items that, read at face value, suggest chaotic behavior (i.e., "Find myself in the same kinds of trouble, time after time," "Am self-destructive," and "Don't know why I do some of the things I do"). Could these items be part of a larger pattern of thoughts, feelings, and behaviors, and would this pattern substantially predict the likelihood of a visit to an emergency room? There's only one way to find out: Broaden the item pool.

3.5.1 Limitations

There are at least two sets of limitations for this study. The first set involves how behavioral frequencies were measured. Behavioral frequencies were self-reported, yearlong patterns that were given a scale from 1 ("Never in my life") to 6 ("More than 20 times in past year"). First, although self-reports of behaviors are valid measures, they are far from perfect; informant ratings are sometimes more valid and often have incremental validity beyond self-reports (Vazire and Mehl, 2008). Second, it was unknown whether a yearlong period for behavioral frequencies was the optimal length of time for the prediction of the selected criteria. Third, the year-to-year stability of yearlong behavioral frequencies was also unknown. Fourth, there was loss of information in the measurement of behavioral frequencies due to participants being limited to six options. All four of these issues are not isolated to the current study; these are limitations of the behavioral frequency literature. The promising results of this study justify further scientific effort to improve the measurement of behavioral frequencies beyond what has been historically acceptable. Future studies should consider: (a) measuring behavior with informant-ratings and/or objective measures; (b) exploring whether different measurement periods for behavioral frequencies may be optimal for different outcomes; (c) determining the stability of behavioral frequencies; and (d) measuring behavioral frequencies with a frequency scale.

The second set of limitations involves the generalizability of this study's results in light of its methods and criteria. BISCUIT may, on average, select fewer items than other statistical learning techniques, and an MMCAR data structure may also influence SLTs toward more parsimonious solutions (Elleman et al., in press; also see Chapter 2). A more complete data structure, or another SLT, such as the elastic net (Zou and Hastie, 2005), could potentially impact (a) the extent to which behavioral frequencies provide unique explanatory variance over traditional personality items, or (b) the distribution of types of personality items in an

SLT's model. Additionally, the small number of criteria in this study should not be assumed to be representative of the much broader pool of life outcomes that researchers may be interested in. Contrary to our position that typically there is congruence between the type of criterion and the type of personality items that best predict it, most criteria of interest may be best predicted by a relatively even mix of cognitions, emotions, and behaviors.

3.5.2 Future Directions

For generalizable findings, studying the incremental validity of self- or informant-reported behavioral frequencies requires a large pool of personality items. Since responding to all items in such a pool would fatigue the average participant, administering a random sample of items is the obvious approach. Thus, researchers may have to rely on MMCAR data structures to further investigate the incremental validity of behavioral frequencies. Immediate next steps include increasing the number and breadth of criteria, refining but also expanding the behavioral frequency item pool, and measuring behaviors on a precise frequency scale.

Although self- and informant-reports have external validity, objective measures are the gold standard. One approach for capturing objective behavioral data is to equip a participant with an always-on Electronically Activated Recorder (EAR; Mehl et al., 2001), which can take the form of a dedicated recording device or be incorporated into a smartphone with a specialized software application (Harari et al., 2016). A major challenge of an EAR-type study is the huge quantity of raw data to be coded, and it would not be feasible for humans to code these data if the time frame were one year. To some extent, machine/statistical learning methods can be used to code phone data, such as summarizing GPS location to identify when a participant has visited a grocery store (Harari et al., 2016; Stachl et al., 2020). However, SLTs will need to advance before they are able to perform tasks akin to text analysis (Iliev et al., 2014; Chen and Wojcik, 2016), turning thousands of hours of video

and audio into frequency variables of how often someone has meditated, slapped someone, or had a hangover in the past year. Coding certain compound behaviors, such as "ate too much," may elude SLTs for some time.

Another promising source of objective behavioral frequency data can be found in the digital footprints that most people make every day. For example, Facebook likes, emails, and credit card transactions can be coded as behaviors in themselves. And just as a clean room leaves behind the behavioral residue of a conscientious individual (Gosling et al., 2002), digital behavioral residue, such as average email response time or the number of minutes spent on a social media app, can be used to infer other behaviors, cognitions, and emotions of interest (Hinds and Joinson, 2019). The primary benefits of studying digital footprints are that the data are objective and already exist in enormous quantities. The primary downside is that the questions that researchers can ask are limited by the data that are available.

3.5.3 Conclusions

Personality is commonly considered to be made up of patterns of cognitions, emotions, and behaviors. The field of personality psychology, however, has not sufficiently investigated behaviors. Self-reported behavioral frequencies are an efficient method of collecting behavioral data on participants by asking raters who are intimate with those behaviors—the participants themselves. The current study presented evidence that behavioral frequencies have incremental validity beyond traditional personality items in describing and accounting for personality-criterion relationships. In terms of the effect sizes of those relationships, in some cases behaviors alone were as good as or even better than cognitions and emotions. The current study found these results using a statistical learning technique, BISCUIT. These results suggest that personality researchers would benefit from expanding the measurement of personality beyond traditional personality items and including more advanced methods

like SLTs in their data analyses. Only by continuing to advance beyond traditional methods and measures may psychologists hope to one day fully lay bare the intricacies of personality.

Chapter 4

General Discussion

The most radical idea of this dissertation is that behavioral frequencies may be thought of as personality items. Many personality psychologists have been resistant to this idea whenever I have discussed it with them. I have not spoken with any researchers who have claimed that patterns of behavior are not part of personality, so I believe this hesitance may be due to personality psychology being almost inextricably tied to traits. Correlating patterns of past behaviors with criteria (which are themselves sometimes behaviors) may seem like circular reasoning. And perhaps more importantly, none of the predictor behaviors point to an *explanation* of what *caused* all of them in the first place.

Personality psychology has been mostly concerned with explaining causes (Yarkoni and Westfall, 2017), and latent traits tend to be the causal explanations for behaviors. Traditional personality items (e.g., "I am a talkative person") are thought to be indicators of latent traits (e.g., Extraversion), and these latent traits are thought to be the causes of behaviors (e.g., talking). Most personality psychologists are comfortable with this explanation. However, this reasoning is just as circular as correlating behaviors with behaviors. Personality psychologists may be more comfortable with this explanation because the circularity is obscured by the aggregation of traditional personality items into traits, often broad domains like the Big

Five. But if the traditional personality item "I am a talkative person" is part of the domain of Extraversion, should we really be impressed that extraverts tend to talk more than introverts? Instances of circularity in personality psychology cannot be fixed by hiding them behind broad domains or trait descriptors, but they can be identified and removed with more surgical measures.

Circularity aside, it is undoubtedly true that traditional personality items, and traditional personality scales such as the Big Five, are valid measures of personality; they predict what one would expect measures of personality to predict (Ozer and Benet-Martínez, 2006; Roberts et al., 2007). To say that traditional personality items measure latent traits, however, is to make an unfalsifiable claim. To use a turn of phrase from Block (1989), the indisputable fact remains that nowhere have traits been directly observed. And because they can never be observed, they can never be falsified. What psychology researchers observe every day are cognitions, emotions, and behaviors. Any theory concerning how these patterns of personality covary is perhaps falsifiable,[1] but the existence of an inner trait is not. Traits do not need to be falsifiable, however, if they are considered to be convenient fictions that summarize the covariances of personality items.

What do personality items measure, if not latent traits? I suspect that most personality psychologists would agree that personality items at least measure patterns of cognitions, emotions, and behaviors. Traditional personality items require that participants summarize these patterns by agreeing or disagreeing with trait descriptors, which are often cognitive or emotional, but sometimes behavioral. If the goal of psychometrics is, as Loevinger (1957) stated, "to isolate, to identify and... to measure separately the important components of

[1]Factor structures probably are not truly falsifiable. As an example, the debate between the Big Five and the HEXACO can never be laid to rest by a preponderance of evidence because both are reasonable solutions and useful in different circumstances, but neither arrives at a higher truth (Srivastava, 2020; Wiernik et al., 2020). There are too many vested interests, too many opinions about which rotation is most appropriate, and too many fit statistics for a nail ever to be driven into the coffin of a halfway-tolerable factor solution. If the past accurately predicts the future, the popular factor structures will survive, while the more fringe but equally reasonable solutions eventually will die with their proponents.

variance" (p. 649), then it is worthwhile to pursue the iterative improvement of personality measurement. One avenue of improvement I propose is to add an underutilized measure of personality, the behavioral frequency, to shore up some of the limitations of traditional personality items. The other proposed avenue is the use of statistical learning approaches (especially BISCUIT), which quantify personality-criterion relationships at the level of personality items.

As an algorithm, BISCUIT is an actuarial method that finds a subset of items that most highly correlate with a criterion. It should not be surprising that this kind of approach can be superior in prediction to a method that is not as rigorously empirical; we have known about the advantages of actuarial judgments for 70 years (Kelly and Fiske, 1950; Meehl, 1954; Dawes et al., 1989). Such a method could be labeled atheoretical, empirical, or perhaps godless. But while BISCUIT's quantitative goal of prediction may reek of dust bowl empiricism, it also has a qualitative goal: to be a transparent statistical learning technique that describes personality-criterion relationships more precisely than at the domain or facet level. I hope that this transparency and precision will help to push some theories of personality in new, fruitful directions. But even if BISCUIT turns out to have limited utility as a theory generator, it is a useful statistical learning technique that finds a balance between prediction and parsimony.

Future Directions: Big Data

As personality researchers continue to delve into the unfathomable depths of Big Data, they will need more tools like BISCUIT; compared to even the largest item pools of a typical personality study, Big Data is a massively missing item ocean. Using empirical approaches like BISCUIT can result in powerful effects in the real world. For example, in one of the first studies to combine personality psychology, Big Data, and field experimentation, Matz et al. (2017) targeted participants with a personality-congruent advertisement and nearly doubled

the purchase rates for an online product, compared to purchases based on a personality-incongruent ad. To achieve this effect, a similar method to BISCUIT was used to determine the top Facebook likes that were correlated with the Extraversion domain, from an item pool of over 65,000 Facebook Likes. These top Likes functioned as an empirical scale of Extraversion, this scale scored millions of new participants, and these scores determined ad congruence.

This study and others that have paired the Big Five with digital footprints (e.g., Kosinski et al., 2013; Gladstone et al., 2019) should be thought of as proofs of concept, not blueprints for endless future iterations. Previous research has indicated that summarizing one hundred self-report trait descriptors into five broad domains results in substantial loss of predictive accuracy. Summarizing petabytes of data in the same way would be catastrophically wasteful. To move forward, personality psychologists will need to reevaluate what they consider to be a personality item. Incorporating self-reports of yearlong patterns of behaviors is a good first step. However, a broader definition of "behavioral frequencies" would include every keystroke, click, transaction, post, and reaction from a person's winding path of digital footprints.

In terms of real-world utility, personality psychology has been a vastly undervalued and underutilized field of study. Tech companies, advertisement firms, and political campaigns are just beginning to understand what personality psychologists and psychometricians have been publishing for over a century: that each person has stable patterns of cognitions, emotions, and behaviors, and that these patterns can be used to predict and influence future cognitions, emotions, and behaviors. Big Data contains an unprecedented wealth of non-traditional personality items disguised as Tweets, Likes, and purchases. If personality psychology evolves into a discipline that fully harnesses the potential of Big Data, it may prove to be the most influential social science of the 21st century. But first, the field needs to move away from traditional conceptions of personality, toward more advanced approaches in analysis and measurement.

References

Allik, J., Realo, A., and McCrae, R. R. (2012). Universality of the five-factor model of personality. In Widiger, T. A., and Costa, P. T. Jr. (Eds.), *Personality disorders and the five-factor model of personality* (3rd ed., pp. 61–74). American Psychological Association, Washington. doi:10.1037/13939-005.

Allport, G. W. and Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1):i–171. doi:10.1037/h0093360.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*:40–79. doi:10.1214/09-SS054.

Asendorpf, J. B., Baumert, A., Schmitt, M., Blum, G., van Bork, R., Rhemtulla, M., Borsboom, D., Chapman, B. P., Clark, D. A., Durbin, C. E., Hicks, B. M., Condon, D. M., Mroczek, D. K., Costantini, G., Perugini, M., Freese, J., Goldberg, L. R., McCrae, R. R., Nave, C. S., ... Mõttus, R. (2016). Open Peer Commentary and Author's Response. *European Journal of Personality*, *30*(4):304–340. doi:10.1002/per.2060.

Ashton, M. C., Lee, K., and Goldberg, L. R. (2007). The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model. *Personality and Individual Differences*, *42*(8):1515–1526. doi:10.1016/j.paid.2006.10.027.

Audigier, V., Husson, F., and Josse, J. (2014). Multiple imputation for continuous variables

using a Bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, *86*(11):2140–2156. doi:10.1080/00949655.2015.1104683.

Audigier, V., Husson, F., and Josse, J. (2016). A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, *10*(1):5–26. doi:10.1007/s11634-014-0195-1.

Bashaw, W. L. and Anderson, H. E. (1967). A Correction for Replicated Error in Correlation Coefficients. *Psychometrika*, *32*(4):435–441. doi:10.1007/BF02289657.

Baumeister, R. F., Vohs, K. D., and Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? *Perspectives on Psychological Science*, *2*(4):396–403. doi:10.1111/j.1745-6916.2007.00051.x.

Block, J. (1989). Critique of the act frequency approach to personality. *Journal of Personality and Social Psychology*, *56*(2):234–245. doi:10.1037/0022-3514.56.2.234.

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, *24*(2):123–140. doi:10.1007/BF00058655.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1):5–32. doi:10.1023/A:1010933404324.

Buss, D. M. and Craik, K. H. (1980). The frequency concept of disposition: dominance and prototypically dominant acts. *Journal of Personality*, *48*(3):379–392. doi:10.1111/j.1467-6494.1980.tb00840.x.

Buss, D. M. and Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*(2):105–126. doi:10.1037/0033-295x.90.2.105.

Buss, D. M. and Craik, K. H. (1985). Why not measure that trait? Alternative criteria for identifying important dispositions. *Journal of Personality and Social Psychology*, *48*(4):934–946. doi:10.1037/0022-3514.48.4.934.

Buss, D. M. and Craik, K. H. (1987). Act Criteria for the Diagnosis of Personality Disorders. *Journal of Personality Disorders*, *1*(1):73–81. doi:10.1521/pedi.1987.1.1.73.

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., and Laibson, D. I. (2015). The Fourth Law of Behavior Genetics. *Current Directions in Psychological Science*, *24*(4):304–312. doi:10.1177/0963721415580430.

Chapman, B. P. and Goldberg, L. R. (2017). Act-frequency signatures of the Big Five. *Personality and Individual Differences*, *116*:201–205. doi:10.1016/j.paid.2017.04.049.

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

Chen, E. E. and Wojcik, S. P. (2016). A Practical Guide to Big Data Research in Psychology. *Psychological Methods*, *21*(4):458–474. doi:10.1037/met0000111.

Church, A. T., Katigbak, M. S., Miramontes, L. G., del Prado, A. M., and Cabrera, H. F. (2007). Culture and the behavioural manifestations of traits: an application of the Act Frequency Approach. *European Journal of Personality*, *21*(4):389–417. doi:10.1002/per.631.

Condon, D. M. (2014). *An organizational framework for the psychological individual differences: Integrating the affective, cognitive, and conative domains* (Doctoral Dissertation). Department of Psychology, Northwestern University, Evanston, IL. doi:10.13140/2.1.4964.1283.

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. doi:10.31234/osf.io/sc4p9.

Condon, D. M. and Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3*(1). doi:10.5334/jopd.al.

Condon, D. M., Roney, E., and Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data*, *5*(1). doi:10.5334/jopd.32.

Costa, P. T. and McCrae, R. R. (1985). *The NEO Personality Inventory*. Psychological Assessment Resources, Odessa, FL.

Costa, P. T. and McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five-factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources, Odessa, FL.

Costa, P. T. and McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing* (pp. 179–198). SAGE Publications, London, UK. doi:10.4135/9781849200479.n9.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1):87–114. doi:10.1017/S0140525X01003922.

Credé, M., Tynan, M. C., and Harms, P. D. (2017). Much ado about grit: A meta-analytic

synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3):492–511. doi:10.1037/pspp0000102.

Csikszentmihalyi, M. and Larson, R. (1987). Validity and reliability of the experience-sampling method. *Journal of Nervous and Mental Disease*, *175*(9):526–536. doi:10.1097/00005053-198709000-00004.

Cureton, E. E. (1966). Corrected Item-Test Correlations. *Psychometrika*, *31*(1):93–93. doi:10.1007/BF02289461.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7):571–582. doi:10.1037/0003-066X.34.7.571.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899):1668–1674. doi:10.1126/science.2648573.

DeYoung, C. G. (2010). Toward a Theory of the Big Five. *Psychological Inquiry*, *21*(1):26–33. doi:10.1080/10478401003648674.

DeYoung, C. G., Hirsh, J. B., Shane, M. S., Papademetris, X., Rajeevan, N., and Gray, J. R. (2010). Testing predictions from personality neuroscience. Brain structure and the big five. *Psychological Science*, *21*(6):820–828. doi:10.1177/0956797610370159.

DeYoung, C. G., Quilty, L. C., and Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5):880–896. doi:10.1037/0022-3514.93.5.880.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*:52–64. doi:10.2307/2282330.

Elleman, L. G., Condon, D. M., Holtzman, N. S., Allen, V. R., and Revelle, W. (2020). *Smaller is better: Associations between personality and demographics are improved by examining narrower traits and regions.* Manuscript submitted for publication. doi:10.31234/osf.io/dpmg2.

Elleman, L. G., Condon, D. M., McDougald, S. K., and Revelle, W. (in press). That takes the BISCUIT: Predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. *European Journal of Psychological Assessment*.

Elleman, L. G., Condon, D. M., and Revelle, W. (2017). Behaviors predict outcomes better than the Big Five. Paper presented at the meeting of the Association for Research in Personality, Sacramento, CA. https://lorienelleman.files.wordpress.com/2019/11/arp2017poster.pdf.

Elleman, L. G., Condon, D. M., and Revelle, W. (2018). Behavioral frequencies strengthen the link between personality and health behaviors. Paper presented at the meeting of the European Association for Personality Psychology, Zadar, Croatia. https://lorienelleman.files.wordpress.com/2019/11/lorienellemanecp2018final.pptx.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*(8):773–790. doi:10.1016/0191-8869(91)90144-z.

Eysenck, S. B. G., Eysenck, H. J., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, *6*(1):21–29. doi:10.1016/0191-8869(85)90026-1.

Fiske, D. W. (1949). Consistency of the Factorial Structures of Personality Ratings

From Different Sources. *Journal of Abnormal and Social Psychology*, *44*(3):329–344. doi:10.1037/h0057198.

Friedman, J. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1):1–67. doi:10.1214/aos/1176347963.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1):1–22. http://www.jstatsoft.org/v33/i01/.

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*(1):197–221. doi:10.1146/annurev.psych.52.1.197.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, *23*(5):369–401. doi:10.1002/per.724.

Gladstone, J. J., Matz, S. C., and Lemaire, A. (2019). Can Psychological Traits Be Inferred From Spending? Evidence From Transaction Data. *Psychological Science*, *30*(7):1087–1096. doi:10.1177/0956797619849435.

Goldberg, L. R. (1971). A historical survey of personality scales and inventories. In McReynolds, P. (Ed.), *Advances in psychological assessment: Volume 2* (pp. 293–336). Science and Behavior Books, Palo Alto, CA.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*(6):1216–1229. doi:10.1037/0022-3514.59.6.1216.

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In Funder, D. C., Parke, R. D., Tomlinson-Keasey, C., and Widaman, K. (Eds.), *Studying*

lives through time: Personality and development (pp. 169–188). American Psychological Association, Washington, DC. doi:10.1037/10127-024.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In Mervielde, I., Deary, I., De Fruyt, F., and Ostendorf, F. (Eds.), *Personality psychology in Europe* (pp. 7–28). Tilburg University Press., Tilburg, Netherlands.

Goldberg, L. R. (2010). Personality, Demographics, and Self-Reported Behavioral Acts: The Development of Avocational Interest Scales from Estimates of the Amount of Time Spent in Interest-Related Activities. In *Then A Miracle Occurs: Focusing on Behavior in Social Psychological Theory and Research* (pp. 205–226). Oxford University Press. doi:10.1093/acprof:oso/9780195377798.003.0011.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1):84–96. doi:10.1016/j.jrp.2005.08.007.

Goldberg, L. R. and Saucier, G. (2016). *The Eugene-Springfield Community Sample: Information Available from the Research Participants*. Technical report. https://ipip.ori.org/ESCS_TechnicalReport_January2016.pdf.

Gosling, S. D., John, O. P., Craik, K. H., and Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology*, *74*(5):1337–1349. doi:10.1037/0022-3514.74.5.1337.

Gosling, S. D., Ko, S. J., Mannarelli, T., and 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. doi:10.1037/0022-3514.82.3.379.

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

Grucza, R. A. and Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment*, *89*(2):167–187. doi:10.1080/00223890701468568.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., and Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science. *Perspectives on Psychological Science*, *11*(6):838–854. doi:10.1177/1745691616650285.

Hathaway, S. R. and McKinley, J. C. (1942). *Manual for the Minnesota Multiphasic Personality Inventory*. University of Minnesota Press, Minneapolis, MN.

Hinds, J. and Joinson, A. (2019). Human and Computer Personality Prediction From Digital Footprints. *Current Directions in Psychological Science*, *28*(2):204–211. doi:10.1177/0963721419827849.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2):95–108. doi:10.1038/nrg1521.

Hirsh, J. B., DeYoung, C. G., and Peterson, J. B. (2009). Metatraits of the Big Five Differentially Predict Engagement and Restraint of Behavior. *Journal of Personality*, *77*(4):1085–1102. doi:10.1111/j.1467-6494.2009.00575.x.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, *1*:278–282. Montreal, Quebec, Canada. doi:10.1109/ICDAR.1995.598994.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1):55–67. doi:10.1080/00401706.1970.10488634.

Hofstee, W. K., de Raad, B., and Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1):146–163. doi:10.1037/0022-3514.63.1.146..

Hogan, R. and Foster, J. (2016). Rethinking personality. *International Journal of Personality Psychology*, *2*(1):37–43. http://ijpp.rug.nl/article/view/25245.

Hogan, R. and Foster, J. (2017). Two further problems with Trait Theory. *International Journal of Personality Psychology*, *3*(1):23–25. http://ijpp.rug.nl/article/view/28408.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2):65–70. doi:10.2307/4615733.

Hope, A. C. A. (1968). A Simplified Monte Carlo Significance Test Procedure. *Journal of the Royal Statistical Society. Series B. Methodological*, *30*(3):582–598. doi:10.1111/j.2517-6161.1968.tb00759.x.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2):179–185. doi:10.1007/BF02289447.

Iliev, R., Dehghani, M., and Sagi, E. (2014). Automated text analysis in psychology: methods, applications, and future developments. *Language and Cognition*, *7*(2):265–290. doi:10.1017/langcog.2014.30.

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., and Roberts, B. W. (2010). What do conscientious people do? Development and validation of the Behavioral

Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, *44*(4):501–511. doi:10.1016/j.jrp.2010.06.005.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY.

Jang, K. L., McCrae, R. R., Angleitner, A., Riemann, R., and Livesley, W. J. (1998). Heritability of facet-level traits in a cross-cultural twin sample: support for a hierarchical model of personality. *Journal of Personality and Social Psychology*, *74*(6):1556–1565. doi:10.1037/0022-3514.74.6.1556.

Jang, K. L., McCrae, R. R., Angleitner, A., Riemann, R., and Livesley, W. J. (1991). The basic level in personality-trait hierarchies: studies of trait use and accessibility in different contexts. *Journal of Personality and Social Psychology*, *60*(3):348–361. doi:10.1037/0022-3514.60.3.348.

John, O. P., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. In John, O. P., Robins, R. W., and Pervin, L. A. (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 114–158). Guilford Press, New York, NY.

John, O. P. and Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In Pervin, L. A. and John, O. P. (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press, New York, NY.

Josse, J. and Husson, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Societe Francaise de Statistique*, *153*(2):79–99. http://journal-sfds.fr/article/view/122.

Josse, J. and Husson, F. (2016). missMDA: A Package for Handling Missing Val-

ues in Multivariate Data Analysis. *Journal of Statistical Software*, *70*(1):1–31. doi:10.18637/jss.v070.i01.

Kelly, E. L. and Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *The American Psychologist*, *5*(8):395–406. doi:10.1037/h0062436.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805. doi:10.1073/pnas.1218772110.

Lee, K. and Ashton, M. C. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research*, *39*(2):329–358. doi:10.1207/s15327906mbr3902_8.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3):18–22.

Loehlin, J. C. (1998). Heritabilities of Common and Measure-Specific Components of the Big Five Personality Factors. *Journal of Research in Personality*, *32*:431–453. doi:10.1006/jrpe.1998.2225.

Loehlin, J. C. and Nichols, R. C. (1976). *Heredity, Environment, & Personality: A Study of 850 Sets of Twins*. University of Texas Press, Austin, TX.

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3):635–694. doi:10.2466/pr0.1957.3.3.635.

Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, *114*(48):12714–12719. doi:10.1073/pnas.1710966114.

McCrae, R. R. (2015). A More Nuanced View of Reliability: Specificity in the Trait Hierarchy. *Personality and Social Psychology Review*, *19*(2):97–112. doi:10.1177/1088868314541857.

Meaney, M. J. (2010). Epigenetics and the biological definition of gene x environment interactions. *Child Development*, *81*(1):41–79. doi:10.1111/j.1467-8624.2009.01381.x.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press, Minneapolis, MN. doi:10.1037/11281-000.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., and Price, J. H. (2001). The Electronically Activated Recorder (EAR): a device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4):517–523. doi:10.3758/bf03195410.

Mershon, B. and Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, *55*(4):675–680. doi:10.1037/0022-3514.55.4.675.

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2):81–97. doi:10.1037/h0043158.

Moser, K. (1989). The act-frequency approach: A conceptual critique. *Personality and Social Psychology Bulletin*, *15*(1):73–83. doi:10.1177/0146167289151007.

Mõttus, R. (2016). Towards more rigorous personality trait-outcome research. *European Journal of Personality*, *30*(4):292–303. doi:10.1002/per.2041.

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D. K., and Revelle, W. (2017, June 23). Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. doi:10.31234/osf.io/4q9gv.

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., and McCrae, R. R. (2017). Personality Traits Below Facets: The Consensual Validity, Longitudinal Stability, Heritability, and Utility of Personality Nuances. *Journal of Personality and Social Psychology*, *112*(3):474–490. doi:10.1037/pspp0000100.

Mõttus, R., McCrae, R. R., Allik, J., and Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*:47–54. doi:10.1016/j.jrp.2014.07.005.

Mõttus, R., Realo, A., Allik, J., Esko, T., and Metspalu, A. (2012). History of the diagnosis of a sexually transmitted disease is linked to normal variation in personality traits. *The Journal of Sexual Medicine*, *9*(11):2861–2867. doi:10.1111/j.1743-6109.2012.02891.x.

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., and Johnson, W. (2015). Within-Trait Heterogeneity in Age Group Differences in Personality Domains and Facets: Implications for the Development and Coherence of Personality Traits. *PLOS ONE*, *10*(3):e0119667. doi:10.1371/journal.pone.0119667.

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Ando, J., Mortensen, E. L., Colodro-Conde, L., and Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *117*(4):e35–e50. doi:10.1037/pspp0000202.

Ozer, D. J. and Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*(1):401–421. doi:10.1146/annurev.psych.57.102904.190127.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of be-

havior. *Journal of Personality and Social Psychology*, *74*(2):538–556. doi:10.1037/0022-3514.74.2.538.

Paunonen, S. V. and Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*(3):524–539. doi:10.1037/0022-3514.81.3.524.

Paunonen, S. V., Haddock, G., Forsterling, F., and Keinonen, M. (2003). Broad versus narrow personality measures and the prediction of behaviour across cultures. *European Journal of Personality*, *17*(6):413–433. doi:10.1002/per.496.

Pozzebon, J. A., Visser, B. A., Ashton, M. C., Lee, K., and Goldberg, L. R. (2010). Psychometric Characteristics of a Public-Domain Self-Report Measure of Vocational Interests: The Oregon Vocational Interest Scales. *Journal of Personality Assessment*, *92*(2):168–174. doi:10.1080/00223890903510431.

Probst, P. and Boulesteix, A. L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, *18*:1–8. http://jmlr.org/papers/volume18/17-269/17-269.pdf.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rentfrow, P. J. (2014). Geographical differences in personality. In Rentfrow, P. J. (Ed.), *Geographical psychology: Exploring the interaction of environment and behavior*. American Psychological Association, Washington, DC.

Revelle, W. (1983). Factors are fictions, and other comments on individuality theory. *Journal of Personality*, *51*(4): 707–714. doi:10.1111/j.1467-6494.1983.tb00875.x.

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, IL. https://CRAN.R-project.org/package=psych.

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., and Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. In Fielding, N. G., Lee, R. M., and Blank, G. (Eds.), *Handbook of Online Research Methods*. Sage Publications, Thousand Oaks, CA.

Revelle, W., Dworak, E. M., and Condon, D. M. (2020). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, 109905. doi:10.1016/j.paid.2020.109905.

Revelle, W., Wilt, J., and Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In Gruszka A., Matthews G., and Szymura B. (Eds.), *Handbook of Individual Differences in Cognition* (pp. 27–49). Springer New York, New York, NY. doi:10.1007/978-1-4419-1210-7_2.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., and Goldberg, L. R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*, *2*(4):313–345. doi:10.1111/j.1745-6916.2007.00047.x.

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA. http://www.rstudio.com/.

Schmitt, P., Mandel, J., and Guedj, M. (2015). A Comparison of Six Methods for Missing Data Imputation. *Journal of Biometrics & Biostatistics*, *6*(1). doi:10.4172/2155-6180.1000224.

Seeboth, A. and Mõttus, R. (2018). Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions. *European Journal of Personality*, *32*:186–201. doi:10.1002/per.2147.

Skimina, E., Cieciuch, J., Schwartz, S. H., Davidov, E., and Algesheimer, R. (2019). Behavioral Signatures of Values in Everyday Behavior in Retrospective and Real-Time Self-Reports. *Frontiers in Psychology*, *10*:521–24. doi:10.3389/fpsyg.2019.00281.

Soto, C. J. and John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*(1):117–143. doi:10.1037/pspp0000096.

Srivastava, S. (2020, April 21). Personality Structure: Who Cares? doi:10.31234/osf.io/dvb4n.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., and Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 201920484. doi:10.1073/pnas.1920484117.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2):245–251. doi:10.1037/0033-2909.87.2.245.

Tang, C., Garreau, D., and von Luxburg, U. (2018). When do random forests fail? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. Garnett, R. (Eds.) *Advances in Neural Information Processing Systems 31*. https://papers.nips.cc/paper/7562-when-do-random-forests-fail.pdf.

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., Uda, M., and Costa Jr, P. T. (2009). Facets of Personality Linked to Underweight and Overweight. *Psychosomatic Medicine*, *71*(6):682–689. doi:10.1097/PSY.0b013e3181a2925b.

Thalmayer, A. G., Saucier, G., and Eigenhuis, A. (2011). Comparative validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*, *23*(4):995–1009. doi:10.1037/a0024165.

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science, 9*(5), 160–164. doi:10.1111/1467-8721.00084.

Vainik, U., Mõttus, R., Allik, J., Esko, T., and Realo, A. (2015). Are trait-outcome associations caused by scales or particular items? Example analysis of personality facets and BMI. *European Journal of Personality*, *29*(6):622–634. doi:10.1002/per.2009.

Vazire, S. and 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. doi:10.1037/a0013314.

Vukasović, T. and Bratko, D. (2015). Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin*, *141*(4):769–785. doi:10.1037/bul0000017.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2):213–217. doi:10.1037/0033-2909.83.2.213.

Waller, N. G. (2008). Fungible Weights in Multiple Regression. *Psychometrika*, *73*(4):691–703. doi:10.1007/s11336-008-9066-z.

Wiernik, B. M., Yarkoni, T., Giordano, C., and Raghavan, M. (2020, April 22). Two, five, six, eight (thousand): Time to end the dimension reduction debate! *Psychometrika*, *73*(4):691–703. doi:10.31234/osf.io/d7jye.

Wilks, S. S. (1938). Weighting Systems for Linear Functions of Correlated Variables When There Is No Dependent Variable. *Psychometrika*, *3*(1):23–40. doi:10.1007/BF02287917.

Wilt, J. (2014). *A New Form and Function for Personality* (Doctoral Dissertation). Department of Psychology, Northwestern University, Evanston, IL. doi:10.1037/e571452013-180.

Wilt, J. and Revelle, W. (2015). Affect, behaviour, cognition and desire in the Big Five: An analysis of item content and structure. *European Journal of Personality*, *29*(4):478–497. doi:10.1002/per.2002.

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

Zou, H. and 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. doi:10.1111/j.1467-9868.2005.00503.x.

Appendix A

Appendix for Chapter 2

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \left[\lambda \sum_{j=1}^{p}|\beta_j|\right] \tag{A.1}$$

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \left[\lambda \sum_{j=1}^{p}\beta_j^2\right] \tag{A.2}$$

Equations for the lasso (equation A.1) and ridge regression (equation A.2). RSS is the residual sum of squares. For both equations, the first term is the typical least squares solution, and the bracketed term is the shrinkage penalty. For the lasso, the penalty is determined by the magnitude of its tuning parameter λ and the absolute value of the β coefficients. For ridge regression, the penalty is determined by the magnitude of its tuning parameter λ and the square of the β coefficients.

Table A.1: Descriptive statistics of participant variables, for the initial sample (no restrictions) and the final sample (complete data for personality items and criteria).

	Initial Sample	Final Sample
Sample size	497,048	78,828
Median age (years)	26	33
Age median absolute deviation	12	19
Percent female	64	65
Number of countries represented	228	200
Percent from United States	39	57
U.S percent Caucasian American	78	81
U.S percent Hispanic American	7	6
U.S percent Asian American	4	4
U.S percent African American	4	3
U.S percent Native American	1	1
U.S percent "other" American	6	6

Table A.2: Average pairwise administrations for training and test subsamples, by item pair type, for all data missingness conditions. Standard deviations are in parentheses. Average item-to-criterion pairwise administrations are larger than item-to-item administrations because criterion data were complete in all conditions.

Subsample	Pair Type	90% missing	75% missing	50% missing	25% missing	Complete data
Training	Item-to-criterion	5,693(69)	14,890(117)	29,342(105)	44,231(105)	59,121(0)
	Item-to-item	510(23)	3,667(61)	14,452(95)	33,008(120)	59,121(0)
Test	Item-to-criterion	1,898(42)	4,963(57)	9,781(79)	14,744(62)	19,707(0)
	Item-to-item	170(13)	1,222(33)	4,817(65)	11,003(71)	19,707(0)

Table A.3: Number of items selected by **BISCUIT,** for each criterion and level of personality data missingness.

Data missingness	Sleep quality	BMI	General health	Education	Smoking frequency
90% missing	1	1	1	81	1
75% missing	1	1	50	69	50
50% missing	42	40	45	38	27
25% missing	30	39	27	38	25
Complete data	6	40	27	37	21

Table A.4: Number of items selected by **the lasso,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations.

Data missingness	Sleep quality	BMI	General health	Education	Smoking frequency
90% missing	26	22	35	27	14
75% missing	40	56	51	82	31
50% missing	42	71	61	101	38
25% missing	48	81	76	112	61
Complete data	56	80	78	108	70

Table A.5: Number of items selected by **the elastic net,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations.

Data missingness	Sleep quality	BMI	General health	Education	Smoking frequency
90% missing	29	21	34	36	15
75% missing	42	58	54	84	32
50% missing	44	73	63	102	41
25% missing	50	81	75	113	62
Complete data	58	78	78	112	77

Table A.6: The six-item model selected by the BISCUIT to predict **Sleep Quality in the complete data condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion.

Item	Mean r	SD r
Am happy with my life.	.33	.001
Dislike myself.	−.32	.001
Feel a sense of worthlessness or hopelessness.	−.32	.001
Feel comfortable with myself.	.29	.001
Love life.	.29	.001
Worry about things.	−.24	.001

Figure A.1: Predictive accuracy (measured in R^2) of the six statistical techniques, using personality data, across five levels of imposed data missingness, in five criteria.

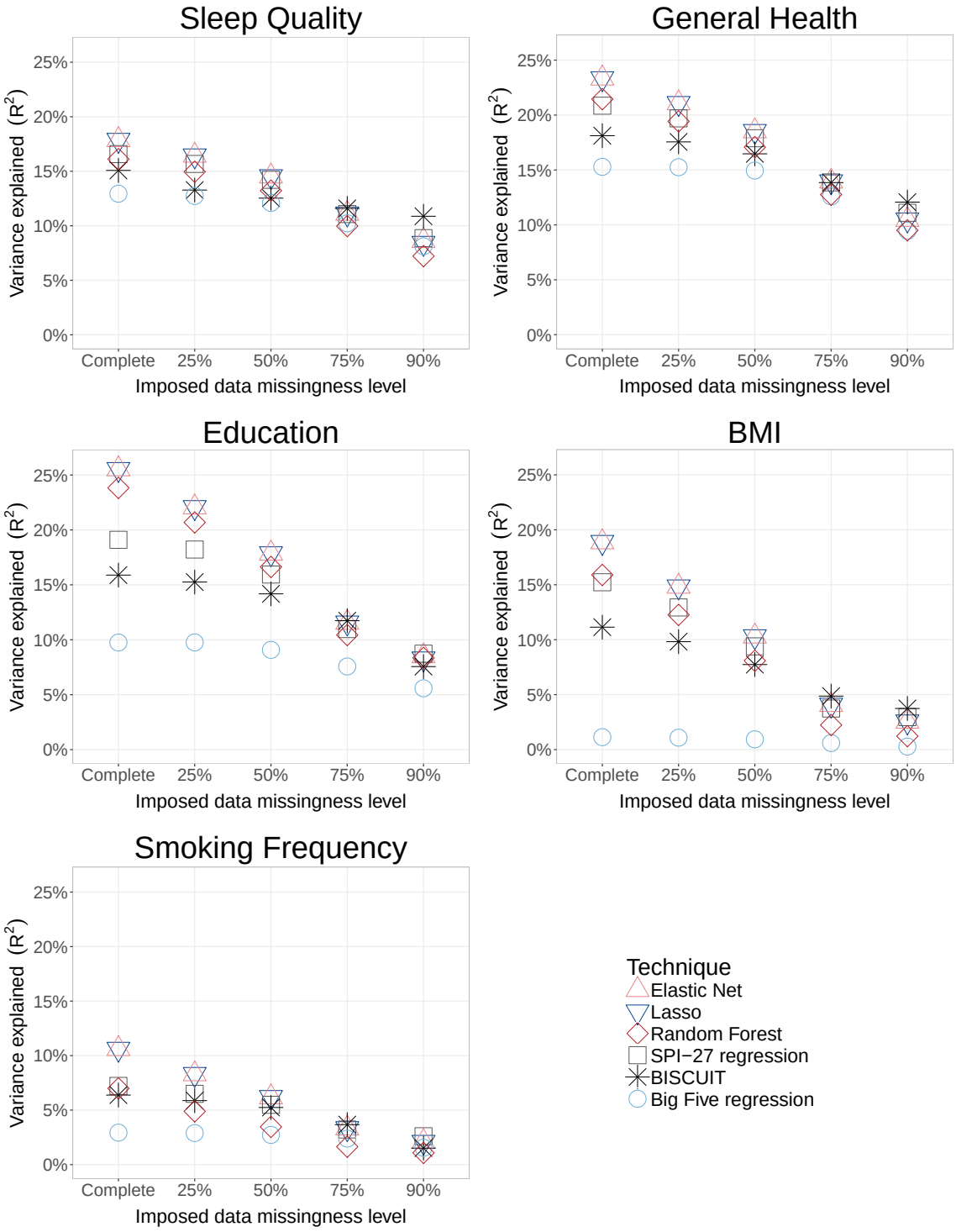


Table A.7: The one-item model selected by the BISCUIT to predict **BMI in the 90% data missingness condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion.

Item	Mean r	SD r
Am able to control my cravings.	−.24	.006

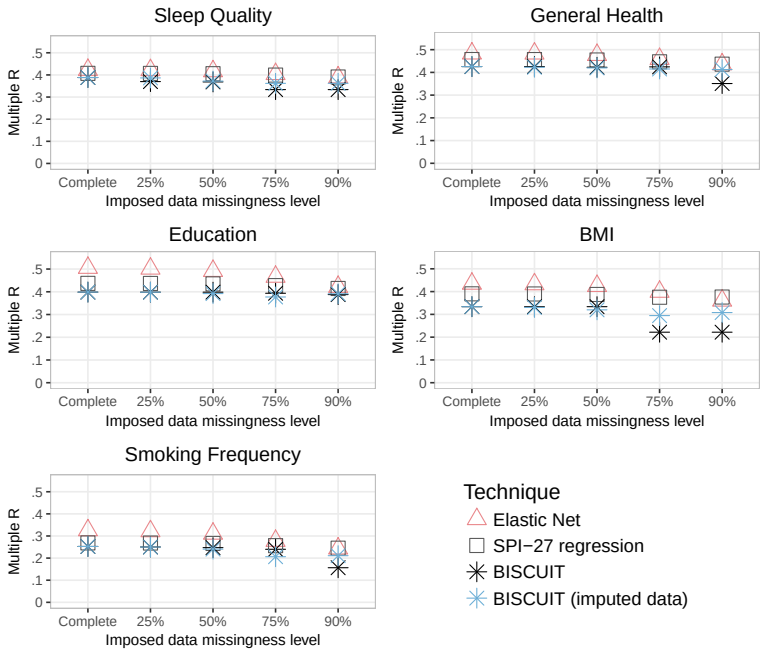
Table A.8: The one-item model selected by the BISCUIT to predict **General Health in the 90% data missingness condition.** BISCUIT unit-weights items in its model, but output from the BISCUIT model includes information regarding the mean and SD of correlations (across folds) of items with the criterion.

Item	Mean r	SD r
Am happy with my life.	.35	.004

Table A.9: Number of items selected by **BISCUIT run on imputed data,** for each criterion and level of personality data missingness. For the 25%, 50%, and 75% data missingness conditions, the number of items is an average across 20 imputations. Mean number of items = 39; median = 37.

Data missingness	Sleep quality	BMI	General health	Education	Smoking frequency
90% missing	43	24	77	37	101
75% missing	40	23	49	53	97
50% missing	28	35	40	39	26
25% missing	7	35	27	38	25
Complete data	6	40	27	37	21

Figure A.2: Predictive accuracy (measured in multiple R) of four techniques based on personality data, across five levels of imposed missingness of data, in five criteria. **Models were only trained on data with an imposed data missingness level; the predictive accuracy of each technique was tested on complete data.**



When tested on complete data, BISCUIT no longer had the highest predictive accuracy in any data missingness condition (Figure A.2). We interpreted this as follows: (a) BISCUIT had selected a number of one-item models due to high levels of data missingness. (b) BISCUIT's simple models cross-validated well when they were tested on data sets with high levels of data missingness. (c) When complete data were used to test all models, however, more complex models had higher predictive accuracy because the complete data provided a better signal with which to validate them. We suspected that if BISCUIT were to be run on imputed data, it would select more items and thus would have better predictive accuracy in a complete test data set, compared to BISCUIT run on missing data. We tested this hypothesis and confirmed it; compared to BISCUIT run on data sets with data missingness, BISCUIT run on imputed data sets selected more items (Table A.9) and had greater predictive accuracy at higher levels of data missingness, when tested on complete data (Figure A.2; Table A.16).

Table A.10: Predictive accuracy (measured in multiple R) of **the lasso and elastic net,** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
Lasso	Sleep quality	90% missing	.29
Lasso	Sleep quality	75% missing	.33
Lasso	Sleep quality	50% missing	.38
Lasso	Sleep quality	25% missing	.41
Lasso	Sleep quality	Complete data	.42
Lasso	BMI	90% missing	.16
Lasso	BMI	75% missing	.20
Lasso	BMI	50% missing	.32
Lasso	BMI	25% missing	.39
Lasso	BMI	Complete data	.43
Lasso	General health	90% missing	.32
Lasso	General health	75% missing	.37
Lasso	General health	50% missing	.43
Lasso	General health	25% missing	.46
Lasso	General health	Complete data	.48
Lasso	Education	90% missing	.29
Lasso	Education	75% missing	.34
Lasso	Education	50% missing	.42
Lasso	Education	25% missing	.47
Lasso	Education	Complete data	.51
Lasso	Smoking frequency	90% missing	.14
Lasso	Smoking frequency	75% missing	.18
Lasso	Smoking frequency	50% missing	.25
Lasso	Smoking frequency	25% missing	.29
Lasso	Smoking frequency	Complete data	.33
Elastic Net	Sleep quality	90% missing	.29
Elastic Net	Sleep quality	75% missing	.33
Elastic Net	Sleep quality	50% missing	.38
Elastic Net	Sleep quality	25% missing	.41
Elastic Net	Sleep quality	Complete data	.42
Elastic Net	BMI	90% missing	.16
Elastic Net	BMI	75% missing	.20
Elastic Net	BMI	50% missing	.32
Elastic Net	BMI	25% missing	.39
Elastic Net	BMI	Complete data	.43
Elastic Net	General health	90% missing	.32
Elastic Net	General health	75% missing	.37
Elastic Net	General health	50% missing	.43
Elastic Net	General health	25% missing	.46
Elastic Net	General health	Complete data	.48
Elastic Net	Education	90% missing	.29
Elastic Net	Education	75% missing	.34
Elastic Net	Education	50% missing	.42
Elastic Net	Education	25% missing	.47
Elastic Net	Education	Complete data	.51
Elastic Net	Smoking frequency	90% missing	.14
Elastic Net	Smoking frequency	75% missing	.18
Elastic Net	Smoking frequency	50% missing	.25
Elastic Net	Smoking frequency	25% missing	.29
Elastic Net	Smoking frequency	Complete data	.33

Table A.11: Predictive accuracy (measured in multiple R) of **BISCUIT and the random forest,** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
BISCUIT	Sleep quality	90% missing	.33
BISCUIT	Sleep quality	75% missing	.34
BISCUIT	Sleep quality	50% missing	.35
BISCUIT	Sleep quality	25% missing	.36
BISCUIT	Sleep quality	Complete data	.39
BISCUIT	BMI	90% missing	.19
BISCUIT	BMI	75% missing	.22
BISCUIT	BMI	50% missing	.28
BISCUIT	BMI	25% missing	.31
BISCUIT	BMI	Complete data	.33
BISCUIT	General health	90% missing	.35
BISCUIT	General health	75% missing	.37
BISCUIT	General health	50% missing	.41
BISCUIT	General health	25% missing	.42
BISCUIT	General health	Complete data	.43
BISCUIT	Education	90% missing	.27
BISCUIT	Education	75% missing	.34
BISCUIT	Education	50% missing	.38
BISCUIT	Education	25% missing	.39
BISCUIT	Education	Complete data	.40
BISCUIT	Smoking frequency	90% missing	.12
BISCUIT	Smoking frequency	75% missing	.19
BISCUIT	Smoking frequency	50% missing	.23
BISCUIT	Smoking frequency	25% missing	.24
BISCUIT	Smoking frequency	Complete data	.25
Random Forest	Sleep quality	90% missing	.27
Random Forest	Sleep quality	75% missing	.32
Random Forest	Sleep quality	50% missing	.36
Random Forest	Sleep quality	25% missing	.39
Random Forest	Sleep quality	Complete data	.40
Random Forest	BMI	90% missing	.11
Random Forest	BMI	75% missing	.15
Random Forest	BMI	50% missing	.28
Random Forest	BMI	25% missing	.35
Random Forest	BMI	Complete data	.40
Random Forest	General health	90% missing	.31
Random Forest	General health	75% missing	.36
Random Forest	General health	50% missing	.41
Random Forest	General health	25% missing	.44
Random Forest	General health	Complete data	.46
Random Forest	Education	90% missing	.29
Random Forest	Education	75% missing	.32
Random Forest	Education	50% missing	.41
Random Forest	Education	25% missing	.45
Random Forest	Education	Complete data	.49
Random Forest	Smoking frequency	90% missing	.10
Random Forest	Smoking frequency	75% missing	.13
Random Forest	Smoking frequency	50% missing	.19
Random Forest	Smoking frequency	25% missing	.22
Random Forest	Smoking frequency	Complete data	.26

Table A.12: Predictive accuracy (measured in multiple R) of **regression using the SPI-27 and regression using the Big Five,** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
SPI-27 regression	Sleep quality	90% missing	.30
SPI-27 regression	Sleep quality	75% missing	.33
SPI-27 regression	Sleep quality	50% missing	.38
SPI-27 regression	Sleep quality	25% missing	.40
SPI-27 regression	Sleep quality	Complete data	.41
SPI-27 regression	BMI	90% missing	.17
SPI-27 regression	BMI	75% missing	.19
SPI-27 regression	BMI	50% missing	.31
SPI-27 regression	BMI	25% missing	.36
SPI-27 regression	BMI	Complete data	.39
SPI-27 regression	General health	90% missing	.33
SPI-27 regression	General health	75% missing	.37
SPI-27 regression	General health	50% missing	.42
SPI-27 regression	General health	25% missing	.44
SPI-27 regression	General health	Complete data	.46
SPI-27 regression	Education	90% missing	.30
SPI-27 regression	Education	75% missing	.33
SPI-27 regression	Education	50% missing	.40
SPI-27 regression	Education	25% missing	.43
SPI-27 regression	Education	Complete data	.44
SPI-27 regression	Smoking frequency	90% missing	.16
SPI-27 regression	Smoking frequency	75% missing	.18
SPI-27 regression	Smoking frequency	50% missing	.23
SPI-27 regression	Smoking frequency	25% missing	.25
SPI-27 regression	Smoking frequency	Complete data	.27
Big Five regression	Sleep quality	90% missing	.28
Big Five regression	Sleep quality	75% missing	.32
Big Five regression	Sleep quality	50% missing	.35
Big Five regression	Sleep quality	25% missing	.36
Big Five regression	Sleep quality	Complete data	.36
Big Five regression	BMI	90% missing	.05
Big Five regression	BMI	75% missing	.08
Big Five regression	BMI	50% missing	.10
Big Five regression	BMI	25% missing	.10
Big Five regression	BMI	Complete data	.11
Big Five regression	General health	90% missing	.31
Big Five regression	General health	75% missing	.35
Big Five regression	General health	50% missing	.39
Big Five regression	General health	25% missing	.39
Big Five regression	General health	Complete data	.39
Big Five regression	Education	90% missing	.24
Big Five regression	Education	75% missing	.28
Big Five regression	Education	50% missing	.30
Big Five regression	Education	25% missing	.31
Big Five regression	Education	Complete data	.31
Big Five regression	Smoking frequency	90% missing	.12
Big Five regression	Smoking frequency	75% missing	.15
Big Five regression	Smoking frequency	50% missing	.17
Big Five regression	Smoking frequency	25% missing	.17
Big Five regression	Smoking frequency	Complete data	.17

Table A.13: Predictive accuracy (measured in multiple R) of **BISCUIT using weighted coefficients and BISCUIT using imputed data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
BISCUIT weighted coefficients	Sleep quality	90% missing	.10
BISCUIT weighted coefficients	Sleep quality	75% missing	.17
BISCUIT weighted coefficients	Sleep quality	50% missing	.36
BISCUIT weighted coefficients	Sleep quality	25% missing	.37
BISCUIT weighted coefficients	Sleep quality	Complete data	.39
BISCUIT weighted coefficients	BMI	90% missing	.06
BISCUIT weighted coefficients	BMI	75% missing	.11
BISCUIT weighted coefficients	BMI	50% missing	.31
BISCUIT weighted coefficients	BMI	25% missing	.34
BISCUIT weighted coefficients	BMI	Complete data	.36
BISCUIT weighted coefficients	General health	90% missing	.11
BISCUIT weighted coefficients	General health	75% missing	.38
BISCUIT weighted coefficients	General health	50% missing	.41
BISCUIT weighted coefficients	General health	25% missing	.42
BISCUIT weighted coefficients	General health	Complete data	.43
BISCUIT weighted coefficients	Education	90% missing	.30
BISCUIT weighted coefficients	Education	75% missing	.35
BISCUIT weighted coefficients	Education	50% missing	.37
BISCUIT weighted coefficients	Education	25% missing	.39
BISCUIT weighted coefficients	Education	Complete data	.40
BISCUIT weighted coefficients	Smoking frequency	90% missing	.04
BISCUIT weighted coefficients	Smoking frequency	75% missing	.20
BISCUIT weighted coefficients	Smoking frequency	50% missing	.23
BISCUIT weighted coefficients	Smoking frequency	25% missing	.24
BISCUIT weighted coefficients	Smoking frequency	Complete data	.25
BISCUIT imputed data	Sleep quality	90% missing	.29
BISCUIT imputed data	Sleep quality	75% missing	.32
BISCUIT imputed data	Sleep quality	50% missing	.36
BISCUIT imputed data	Sleep quality	25% missing	.38
BISCUIT imputed data	Sleep quality	Complete data	.39
BISCUIT imputed data	BMI	90% missing	.15
BISCUIT imputed data	BMI	75% missing	.17
BISCUIT imputed data	BMI	50% missing	.26
BISCUIT imputed data	BMI	25% missing	.31
BISCUIT imputed data	BMI	Complete data	.33
BISCUIT imputed data	General health	90% missing	.33
BISCUIT imputed data	General health	75% missing	.36
BISCUIT imputed data	General health	50% missing	.40
BISCUIT imputed data	General health	25% missing	.42
BISCUIT imputed data	General health	Complete data	.43
BISCUIT imputed data	Education	90% missing	.29
BISCUIT imputed data	Education	75% missing	.31
BISCUIT imputed data	Education	50% missing	.37
BISCUIT imputed data	Education	25% missing	.39
BISCUIT imputed data	Education	Complete data	.40
BISCUIT imputed data	Smoking frequency	90% missing	.15
BISCUIT imputed data	Smoking frequency	75% missing	.17
BISCUIT imputed data	Smoking frequency	50% missing	.22
BISCUIT imputed data	Smoking frequency	25% missing	.24
BISCUIT imputed data	Smoking frequency	Complete data	.25

Table A.14: Predictive accuracy (measured in multiple R) of **the elastic net and BISCUIT applied to the SPI-27 (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
Elastic Net SPI-27	Sleep quality	90% missing	.30
Elastic Net SPI-27	Sleep quality	75% missing	.33
Elastic Net SPI-27	Sleep quality	50% missing	.38
Elastic Net SPI-27	Sleep quality	25% missing	.40
Elastic Net SPI-27	Sleep quality	Complete data	.41
Elastic Net SPI-27	BMI	90% missing	.17
Elastic Net SPI-27	BMI	75% missing	.19
Elastic Net SPI-27	BMI	50% missing	.31
Elastic Net SPI-27	BMI	25% missing	.36
Elastic Net SPI-27	BMI	Complete data	.39
Elastic Net SPI-27	General health	90% missing	.33
Elastic Net SPI-27	General health	75% missing	.37
Elastic Net SPI-27	General health	50% missing	.42
Elastic Net SPI-27	General health	25% missing	.44
Elastic Net SPI-27	General health	Complete data	.46
Elastic Net SPI-27	Education	90% missing	.30
Elastic Net SPI-27	Education	75% missing	.33
Elastic Net SPI-27	Education	50% missing	.40
Elastic Net SPI-27	Education	25% missing	.43
Elastic Net SPI-27	Education	Complete data	.44
Elastic Net SPI-27	Smoking frequency	90% missing	.16
Elastic Net SPI-27	Smoking frequency	75% missing	.18
Elastic Net SPI-27	Smoking frequency	50% missing	.23
Elastic Net SPI-27	Smoking frequency	25% missing	.25
Elastic Net SPI-27	Smoking frequency	Complete data	.27
BISCUIT SPI-27	Sleep quality	90% missing	.29
BISCUIT SPI-27	Sleep quality	75% missing	.31
BISCUIT SPI-27	Sleep quality	50% missing	.35
BISCUIT SPI-27	Sleep quality	25% missing	.36
BISCUIT SPI-27	Sleep quality	Complete data	.36
BISCUIT SPI-27	BMI	90% missing	.13
BISCUIT SPI-27	BMI	75% missing	.14
BISCUIT SPI-27	BMI	50% missing	.22
BISCUIT SPI-27	BMI	25% missing	.25
BISCUIT SPI-27	BMI	Complete data	.28
BISCUIT SPI-27	General health	90% missing	.32
BISCUIT SPI-27	General health	75% missing	.36
BISCUIT SPI-27	General health	50% missing	.40
BISCUIT SPI-27	General health	25% missing	.41
BISCUIT SPI-27	General health	Complete data	.42
BISCUIT SPI-27	Education	90% missing	.28
BISCUIT SPI-27	Education	75% missing	.31
BISCUIT SPI-27	Education	50% missing	.35
BISCUIT SPI-27	Education	25% missing	.37
BISCUIT SPI-27	Education	Complete data	.37
BISCUIT SPI-27	Smoking frequency	90% missing	.15
BISCUIT SPI-27	Smoking frequency	75% missing	.17
BISCUIT SPI-27	Smoking frequency	50% missing	.20
BISCUIT SPI-27	Smoking frequency	25% missing	.21
BISCUIT SPI-27	Smoking frequency	Complete data	.21

Table A.15: Predictive accuracy (measured in multiple R) of **the elastic net and SPI-27 trained on imputed data and tested on complete data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
Elastic Net imputed-on-complete	Sleep quality	90% missing	.39
Elastic Net imputed-on-complete	Sleep quality	75% missing	.40
Elastic Net imputed-on-complete	Sleep quality	50% missing	.42
Elastic Net imputed-on-complete	Sleep quality	25% missing	.42
Elastic Net imputed-on-complete	Sleep quality	Complete data	.42
Elastic Net imputed-on-complete	BMI	90% missing	.36
Elastic Net imputed-on-complete	BMI	75% missing	.40
Elastic Net imputed-on-complete	BMI	50% missing	.42
Elastic Net imputed-on-complete	BMI	25% missing	.43
Elastic Net imputed-on-complete	BMI	Complete data	.43
Elastic Net imputed-on-complete	General health	90% missing	.44
Elastic Net imputed-on-complete	General health	75% missing	.46
Elastic Net imputed-on-complete	General health	50% missing	.48
Elastic Net imputed-on-complete	General health	25% missing	.48
Elastic Net imputed-on-complete	General health	Complete data	.48
Elastic Net imputed-on-complete	Education	90% missing	.42
Elastic Net imputed-on-complete	Education	75% missing	.47
Elastic Net imputed-on-complete	Education	50% missing	.49
Elastic Net imputed-on-complete	Education	25% missing	.50
Elastic Net imputed-on-complete	Education	Complete data	.51
Elastic Net imputed-on-complete	Smoking frequency	90% missing	.24
Elastic Net imputed-on-complete	Smoking frequency	75% missing	.28
Elastic Net imputed-on-complete	Smoking frequency	50% missing	.31
Elastic Net imputed-on-complete	Smoking frequency	25% missing	.32
Elastic Net imputed-on-complete	Smoking frequency	Complete data	.33
SPI-27 imputed-on-complete	Sleep quality	90% missing	.39
SPI-27 imputed-on-complete	Sleep quality	75% missing	.40
SPI-27 imputed-on-complete	Sleep quality	50% missing	.41
SPI-27 imputed-on-complete	Sleep quality	25% missing	.41
SPI-27 imputed-on-complete	Sleep quality	Complete data	.41
SPI-27 imputed-on-complete	BMI	90% missing	.38
SPI-27 imputed-on-complete	BMI	75% missing	.38
SPI-27 imputed-on-complete	BMI	50% missing	.39
SPI-27 imputed-on-complete	BMI	25% missing	.39
SPI-27 imputed-on-complete	BMI	Complete data	.39
SPI-27 imputed-on-complete	General health	90% missing	.44
SPI-27 imputed-on-complete	General health	75% missing	.45
SPI-27 imputed-on-complete	General health	50% missing	.45
SPI-27 imputed-on-complete	General health	25% missing	.46
SPI-27 imputed-on-complete	General health	Complete data	.46
SPI-27 imputed-on-complete	Education	90% missing	.41
SPI-27 imputed-on-complete	Education	75% missing	.43
SPI-27 imputed-on-complete	Education	50% missing	.43
SPI-27 imputed-on-complete	Education	25% missing	.44
SPI-27 imputed-on-complete	Education	Complete data	.44
SPI-27 imputed-on-complete	Smoking frequency	90% missing	.24
SPI-27 imputed-on-complete	Smoking frequency	75% missing	.26
SPI-27 imputed-on-complete	Smoking frequency	50% missing	.27
SPI-27 imputed-on-complete	Smoking frequency	25% missing	.27
SPI-27 imputed-on-complete	Smoking frequency	Complete data	.27

Table A.16: Predictive accuracy (measured in multiple R) of **BISCUIT trained on missing data and BISCUIT trained on imputed data, both tested on complete data (post-hoc analyses),** based on personality data, across five levels of imposed missingness of data, in five criteria.

Model	Criterion	Data missingness	Multiple R
BISCUIT missing-on-complete	Sleep quality	90% missing	.33
BISCUIT missing-on-complete	Sleep quality	75% missing	.33
BISCUIT missing-on-complete	Sleep quality	50% missing	.37
BISCUIT missing-on-complete	Sleep quality	25% missing	.37
BISCUIT missing-on-complete	Sleep quality	Complete data	.39
BISCUIT missing-on-complete	BMI	90% missing	.22
BISCUIT missing-on-complete	BMI	75% missing	.22
BISCUIT missing-on-complete	BMI	50% missing	.33
BISCUIT missing-on-complete	BMI	25% missing	.33
BISCUIT missing-on-complete	BMI	Complete data	.33
BISCUIT missing-on-complete	General health	90% missing	.35
BISCUIT missing-on-complete	General health	75% missing	.42
BISCUIT missing-on-complete	General health	50% missing	.42
BISCUIT missing-on-complete	General health	25% missing	.43
BISCUIT missing-on-complete	General health	Complete data	.43
BISCUIT missing-on-complete	Education	90% missing	.39
BISCUIT missing-on-complete	Education	75% missing	.39
BISCUIT missing-on-complete	Education	50% missing	.40
BISCUIT missing-on-complete	Education	25% missing	.40
BISCUIT missing-on-complete	Education	Complete data	.40
BISCUIT missing-on-complete	Smoking frequency	90% missing	.16
BISCUIT missing-on-complete	Smoking frequency	75% missing	.24
BISCUIT missing-on-complete	Smoking frequency	50% missing	.25
BISCUIT missing-on-complete	Smoking frequency	25% missing	.25
BISCUIT missing-on-complete	Smoking frequency	Complete data	.25
BISCUIT imputed-on-complete	Sleep quality	90% missing	.36
BISCUIT imputed-on-complete	Sleep quality	75% missing	.36
BISCUIT imputed-on-complete	Sleep quality	50% missing	.37
BISCUIT imputed-on-complete	Sleep quality	25% missing	.39
BISCUIT imputed-on-complete	Sleep quality	Complete data	.39
BISCUIT imputed-on-complete	BMI	90% missing	.31
BISCUIT imputed-on-complete	BMI	75% missing	.29
BISCUIT imputed-on-complete	BMI	50% missing	.32
BISCUIT imputed-on-complete	BMI	25% missing	.33
BISCUIT imputed-on-complete	BMI	Complete data	.33
BISCUIT imputed-on-complete	General health	90% missing	.41
BISCUIT imputed-on-complete	General health	75% missing	.42
BISCUIT imputed-on-complete	General health	50% missing	.42
BISCUIT imputed-on-complete	General health	25% missing	.42
BISCUIT imputed-on-complete	General health	Complete data	.43
BISCUIT imputed-on-complete	Education	90% missing	.39
BISCUIT imputed-on-complete	Education	75% missing	.38
BISCUIT imputed-on-complete	Education	50% missing	.39
BISCUIT imputed-on-complete	Education	25% missing	.40
BISCUIT imputed-on-complete	Education	Complete data	.40
BISCUIT imputed-on-complete	Smoking frequency	90% missing	.21
BISCUIT imputed-on-complete	Smoking frequency	75% missing	.21
BISCUIT imputed-on-complete	Smoking frequency	50% missing	.24
BISCUIT imputed-on-complete	Smoking frequency	25% missing	.25
BISCUIT imputed-on-complete	Smoking frequency	Complete data	.25

Appendix B

Appendix for Chapter 3

The Creation of the Behavioral Acts, Revised and Expanded (BARE) Inventory

The origin of the BARE began with a pool of 324 behavioral items (the Objective Behavior Inventory) included as part of a twin study (Loehlin and Nichols, 1976). This pool of items was used as the basis for 400 behavioral frequencies given in 1997 to the Eugene-Springfield Community Sample (ESCS), an ongoing longitudinal study (Grucza and Goldberg, 2007; Goldberg and Saucier, 2016). Grucza and Goldberg (2007) stated that the purpose of this item pool was to "develop a reasonably comprehensive pool of activity descriptors" (p. 171). Goldberg (2010) selected a subset of the 400 behaviors and added new ones to create a set of 33 scales composed of 200 items, which he administered to the ESCS in 2007. Goldberg called this new set of scales the Oregon Avocational Interest Scales (ORAIS) and considered it a companion set to the Oregon Vocational Interest Scales (ORVIS; Pozzebon et al., 2010); together, they were to meet the need for public-domain interests scales (Goldberg, 2010). Although there may be some appeal in having a set of both vocational and avocational interest scales, it would be incorrect to call these 200 behavioral frequencies "interests." The format of the ORAIS is identical to that of the earlier 400 behavioral items and to behavioral frequencies in general: participants are asked to report the frequency of their behavior in

a given time frame. Vocational interests, on the other hand, ask participants to rate their level of interest for occupational tasks (e.g., "Be a racing car driver") on a Likert-like scale from "Strongly dislike" to "Strongly like." More recently, Goldberg appears to have agreed that the ORAIS items are in fact behaviors and not avocational interests; Chapman and Goldberg (2017) described the behaviors of the Big Five using the old list of 400 behavioral frequencies, which they called the Behavioral Acts Inventory (BAI).

The SAPA team began administering the ORAIS on May 20, 2013. In order to expand SAPA's behavioral frequency item pool, Lorien Elleman (with the help of Sarah McDougald, an undergraduate research assistant) cross-checked the item content of the ORAIS (available in the appendix of Goldberg, 2010) against the BAI (available in the supplemental materials of Chapman and Goldberg, 2017). Items in the BAI that were duplicates of ORAIS items were eliminated by using keyword matches. This analysis identified 255 behavioral frequencies from the BAI to be added to the 199 unique items of the ORAIS.[1] The wording of 74 of the new items were revised in order to improve their quality, with the goal being to broaden, clarify, specify, or modernize the behaviors. For example, "Spent time preserving or canning fruits or vegetables" was revised to "Spent time preserving, canning, or pickling food"; "Drove more than 200 miles by myself" was revised to "Drove more than 200 miles (325 km) by myself"; "Decorated a room" was revised to "Decorated a room (not as a job)"; and "Rode in a taxi" was revised to "Rode in a taxi or rideshare (such as Lyft or Uber)."

In preparation for publishing an updated inventory of behavioral frequencies, Elleman and McDougald performed three quality control procedures in which each person manually checked the item content of each new behavioral frequency, and the two discussed and agreed upon removing items. The purpose of these procedures was to remove behaviors that were the most glaring examples of items that did not belong in the inventory. A more comprehensive

[1]The 200-item ORAIS is actually 199 unique items. The item "Wrote poetry" is used in two scales: Creativity and Writing/Remembering.

review with a large panel of judges would have been a more optimal approach but was outside the scope and resources of the current study. We are strongly in favor of a project dedicated to the thorough examination of each behavior for inclusion and/or revision in a future iteration of the BARE Inventory.

The first quality control procedure removed six items that were conceptual duplicates of the ORAIS (i.e., duplicates which had not been identified by matching keywords). The second procedure removed three behaviors that were not explicitly performed by the target (e.g., "Was hit or slapped"). The third procedure removed 20 items that lacked face validity of being "activity descriptors" (Grucza and Goldberg, 2007, p. 171) and/or had ambiguity concerning the volition of the behavior, which is a standard that has been used in previous research (Skimina et al., 2019). For example, "Had a stomach ache" could technically be considered a behavior, but not a volitional activity that the target participated in. Combined, these three procedures removed 29 items. Thus, the BARE Inventory consisted of 425 items, 199 of which were from the ORAIS,[2] and 226 of which originated from the BAI.

[2]Due to the 199 items of the ORAIS already having been integrated with the SAPA website, we decided not to revise or remove any items from the ORAIS, and simply include all of them as part of the BARE. Future iterations of the BARE should perhaps revise items which originated from the ORAIS.

Table B.1: Items of the BARE Inventory. "SAPA ID" refers to the unique item identifier used by the SAPA Project. "Origin" refers to whether the item was taken from the ORAIS or the BAI. The last column lists either the ORAIS facet associated with the item or the original BAI item ID number.

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_3322	Ate dinner alone.	ORAIS	Being Alone
q_3323	Chose to spend a day by myself.	ORAIS	Being Alone
q_3324	Spent an entire vacation by myself.	ORAIS	Being Alone
q_3325	Went on a trip by myself.	ORAIS	Being Alone
q_3326	Went to a concert or theater alone.	ORAIS	Being Alone
q_3327	Went to the movies alone.	ORAIS	Being Alone
q_3328	Let a child win a game.	ORAIS	Child-Related
q_3329	Played with a child.	ORAIS	Child-Related
q_3330	Read a story to a child.	ORAIS	Child-Related
q_3331	Read the comics to a child.	ORAIS	Child-Related
q_3332	Served as a baby sitter.	ORAIS	Child-Related
q_3333	Took a child on an outing.	ORAIS	Child-Related
q_3334	Bought a book about the things that I collect.	ORAIS	Collecting
q_3335	Bought something for my collection.	ORAIS	Collecting
q_3336	Read a book about the things that I collect.	ORAIS	Collecting
q_3337	Traded something in my collection.	ORAIS	Collecting
q_3338	Worked on my collection.	ORAIS	Collecting
q_3339	Looked up information on the Internet.	ORAIS	Computing
q_3340	Played a computer game.	ORAIS	Computing
q_3341	Read news on the Internet.	ORAIS	Computing
q_3342	Sent a message by electronic mail (e-mail).	ORAIS	Computing
q_3343	Surfed the Internet.	ORAIS	Computing
q_3344	Used a computer.	ORAIS	Computing
q_3345	Acted in a play.	ORAIS	Creativity
q_3346	Painted a picture.	ORAIS	Creativity
q_3347	Played a musical instrument.	ORAIS	Creativity
q_3348	Produced a work of art.	ORAIS	Creativity
q_3349	Sang or played an instrument in public.	ORAIS	Creativity
q_3350	Tried something completely new.	ORAIS	Creativity
q_3351	Wrote poetry.	ORAIS	Creativity and Writing/Rem.
q_3352	Attended a ballet performance.	ORAIS	Culture
q_3353	Attended a public lecture.	ORAIS	Culture
q_3354	Attended a stage play or musical.	ORAIS	Culture

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_3355	Attended an opera or a concert.	ORAIS	Culture
q_3356	Visited a museum.	ORAIS	Culture
q_3357	Visited an art exhibition.	ORAIS	Culture
q_3358	Became intoxicated.	ORAIS	Drinking
q_3359	Drank beer or wine.	ORAIS	Drinking
q_3360	Drank in a bar or night club.	ORAIS	Drinking
q_3361	Drank whiskey, vodka, gin, or other hard liquor.	ORAIS	Drinking
q_3362	Had a hangover.	ORAIS	Drinking
q_3363	Did aerobic exercise.	ORAIS	Exercise
q_3364	Did yoga or other movement exercises.	ORAIS	Exercise
q_3365	Exercised for 40 minutes or longer.	ORAIS	Exercise
q_3366	Lifted weights.	ORAIS	Exercise
q_3367	Participated in an exercise program.	ORAIS	Exercise
q_3368	Used an exercise machine.	ORAIS	Exercise
q_3369	Went running or jogging.	ORAIS	Exercise
q_3370	Bought a fashionable item of clothing.	ORAIS	Fashion
q_3371	Read a fashion-related book.	ORAIS	Fashion
q_3372	Read a fashion-related magazine.	ORAIS	Fashion
q_3373	Spent more than 10 minutes thinking about what to wear.	ORAIS	Fashion
q_3374	Spent more than an hour thinking about what to wear.	ORAIS	Fashion
q_3375	Bought or sold real estate.	ORAIS	Financial
q_3376	Bought or sold stocks or bonds.	ORAIS	Financial
q_3377	Obtained stock market prices.	ORAIS	Financial
q_3378	Purchased a commodity as an investment.	ORAIS	Financial
q_3379	Read a book on a financial topic.	ORAIS	Financial
q_3380	Worked on a retirement plan.	ORAIS	Financial
q_3381	Ate candy.	ORAIS	Food-Related
q_3382	Ate food while walking or working.	ORAIS	Food-Related
q_3383	Ate in a restaurant.	ORAIS	Food-Related
q_3384	Ate or drank while driving.	ORAIS	Food-Related
q_3385	Ate too much.	ORAIS	Food-Related
q_3386	Chewed gum.	ORAIS	Food-Related
q_3387	Ordered food to be delivered.	ORAIS	Food-Related
q_3388	Bet money on a sports event.	ORAIS	Gambling
q_3389	Gambled on a slot machine or video poker game.	ORAIS	Gambling

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_3390	Gambled with cards or dice.	ORAIS	Gambling
q_3391	Played bingo for money.	ORAIS	Gambling
q_3392	Purchased a scratch ticket.	ORAIS	Gambling
q_3393	Went to a casino.	ORAIS	Gambling
q_3394	Learned a new board or card game.	ORAIS	Game-Playing
q_3395	Played a board game.	ORAIS	Game-Playing
q_3396	Played cards.	ORAIS	Game-Playing
q_3397	Played chess or checkers.	ORAIS	Game-Playing
q_3398	Worked on a jigsaw puzzle.	ORAIS	Game-Playing
q_3399	Bought or picked flowers.	ORAIS	Gardening
q_3400	Bought plants for a garden or yard.	ORAIS	Gardening
q_3401	Cared for a potted plant.	ORAIS	Gardening
q_3402	Did yard work.	ORAIS	Gardening
q_3403	Gardened.	ORAIS	Gardening
q_3404	Planted or transplanted a plant.	ORAIS	Gardening
q_3405	Changed a habit to have less impact on the environment.	ORAIS	Green. Acts.
q_3406	Composted food scraps or yard waste.	ORAIS	Green. Acts.
q_3407	Picked up litter.	ORAIS	Green. Acts.
q_3408	Recycled one or more items.	ORAIS	Green. Acts.
q_3409	Used both sides of a piece of paper before discarding it.	ORAIS	Green. Acts.
q_3410	Used public transportation.	ORAIS	Green. Acts.
q_3411	Walked or rode a bicycle to work.	ORAIS	Green. Acts.
q_3412	Baked a cake, pie, cookies, or bread.	ORAIS	Housekeeping
q_3413	Cleaned the house.	ORAIS	Housekeeping
q_3414	Cooked a meal.	ORAIS	Housekeeping
q_3415	Ironed linens or clothes.	ORAIS	Housekeeping
q_3416	Knitted, quilted, sewed, or crocheted.	ORAIS	Housekeeping
q_3417	Made a bed.	ORAIS	Housekeeping
q_3418	Washed dishes.	ORAIS	Housekeeping
q_3419	Downloaded music from the Internet.	ORAIS	Music
q_3420	Listened to music on the radio.	ORAIS	Music
q_3421	Listened to music while working.	ORAIS	Music
q_3422	Purchased a musical album.	ORAIS	Music
q_3423	Read music-related news.	ORAIS	Music
q_3424	Shopped in a music store.	ORAIS	Music
q_3425	Traded music with a friend.	ORAIS	Music
q_3426	Used a portable music player.	ORAIS	Music

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_3427	Entertained six or more people.	ORAIS	Partying
q_3428	Had someone over for dinner.	ORAIS	Partying
q_3429	Planned a party.	ORAIS	Partying
q_3430	Went to a large party.	ORAIS	Partying
q_3431	Went to a small party.	ORAIS	Partying
q_3432	Bathed or groomed a pet animal.	ORAIS	Pets
q_3433	Cared for a pet animal.	ORAIS	Pets
q_3434	Fed a pet animal.	ORAIS	Pets
q_3435	Played with a pet animal.	ORAIS	Pets
q_3436	Purchased a pet animal.	ORAIS	Pets
q_3437	Attended a rally or demonstration.	ORAIS	Political/Org.
q_3438	Attended a town meeting.	ORAIS	Political/Org.
q_3439	Donated money to a political campaign or cause.	ORAIS	Political/Org.
q_3440	Donated money to charity.	ORAIS	Political/Org.
q_3441	Signed a petition.	ORAIS	Political/Org.
q_3442	Volunteered for a club or organization.	ORAIS	Political/Org.
q_3443	Wrote a letter to a newspaper or politician.	ORAIS	Political/Org.
q_3444	Bought a book.	ORAIS	Reading
q_3445	Read a book.	ORAIS	Reading
q_3446	Read an entire book in one sitting.	ORAIS	Reading
q_3447	Read in bed before going to sleep.	ORAIS	Reading
q_3448	Visited a public library.	ORAIS	Reading
q_3449	Attended a church or religious service.	ORAIS	Religious/Spirit.
q_3450	Discussed religion or spirituality.	ORAIS	Religious/Spirit.
q_3451	Gave a blessing at a meal.	ORAIS	Religious/Spirit.
q_3452	Listened to a religious program on the radio or TV.	ORAIS	Religious/Spirit.
q_3453	Prayed (not including blessings at meals).	ORAIS	Religious/Spirit.
q_3454	Read a book about religion or spirituality.	ORAIS	Religious/Spirit.
q_3455	Read the Bible or other sacred text.	ORAIS	Religious/Spirit.
q_3456	Attended a formal dance.	ORAIS	Romance
q_3457	Dined by candle light.	ORAIS	Romance
q_3458	Went dancing.	ORAIS	Romance
q_3459	Went on a date.	ORAIS	Romance
q_3460	Wore formal clothing.	ORAIS	Romance
q_3461	Wrote a love letter.	ORAIS	Romance
q_3462	Bought a self-help book.	ORAIS	Self-Improve.
q_3463	Enrolled in a course of study.	ORAIS	Self-Improve.

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_3464	Learned a new skill.	ORAIS	Self-Improve.
q_3465	Read a self-help book.	ORAIS	Self-Improve.
q_3466	Studied some subject.	ORAIS	Self-Improve.
q_3467	Bought something other than groceries.	ORAIS	Shopping
q_3468	Checked the sales ads in a newspaper.	ORAIS	Shopping
q_3469	Read newspaper ads for non-grocery items.	ORAIS	Shopping
q_3470	Shopped on the web.	ORAIS	Shopping
q_3471	Spent 10 minutes or more in a non-grocery store.	ORAIS	Shopping
q_3472	Spent an hour or more in a non-grocery store.	ORAIS	Shopping
q_3473	Used eBay to buy or sell something.	ORAIS	Shopping
q_3474	Made an entry on my own Facebook page (or similar social networking website).	ORAIS	Social-Ntwrk.
q_3475	Participated in an online discussion group.	ORAIS	Social-Ntwrk.
q_3476	Read someone's personal web page (including Facebook or similar sites).	ORAIS	Social-Ntwrk.
q_3477	Used a computer for social networking.	ORAIS	Social-Ntwrk.
q_3478	Used instant messaging to chat online.	ORAIS	Social-Ntwrk.
q_3479	Attended an athletic event.	ORAIS	Sports
q_3480	Discussed sports.	ORAIS	Sports
q_3481	Played a team sport.	ORAIS	Sports
q_3482	Played basketball.	ORAIS	Sports
q_3483	Played tennis or golf.	ORAIS	Sports
q_3484	Watched a televised sports event.	ORAIS	Sports
q_3485	Walked on a beach.	ORAIS	Summer Acts.
q_3486	Went backpacking or camping.	ORAIS	Summer Acts.
q_3487	Went boating or rafting.	ORAIS	Summer Acts.
q_3488	Went fishing or hunting.	ORAIS	Summer Acts.
q_3489	Went on a hike.	ORAIS	Summer Acts.
q_3490	Went on a picnic.	ORAIS	Summer Acts.
q_3491	Went swimming.	ORAIS	Summer Acts.
q_3492	Stayed in a hotel, motel, or resort.	ORAIS	Travel
q_3493	Took a trip.	ORAIS	Travel
q_3494	Took travel photographs.	ORAIS	Travel
q_3495	Traveled by train or plane.	ORAIS	Travel
q_3496	Went on a cruise or tour.	ORAIS	Travel
q_3497	Went sightseeing.	ORAIS	Travel
q_3498	Recorded a television program.	ORAIS	TV

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_3499	Watched a television reality show.	ORAIS	TV
q_3500	Watched a television soap opera.	ORAIS	TV
q_3501	Watched a television talk show.	ORAIS	TV
q_3502	Watched television news.	ORAIS	TV
q_3503	Watched television.	ORAIS	TV
q_3504	Watched too much television.	ORAIS	TV
q_3505	Looked something up in an encyclopedia (or wikipedia).	ORAIS	Understanding
q_3506	Looked up a word in a dictionary.	ORAIS	Understanding
q_3507	Read a news magazine.	ORAIS	Understanding
q_3508	Read poetry.	ORAIS	Understanding
q_3509	Read the editorial page of a newspaper.	ORAIS	Understanding
q_3510	Watched an educational channel on TV.	ORAIS	Understanding
q_3511	Bought a car, truck, or motorcycle.	ORAIS	Vehicles
q_3512	Raced a car, truck, or motorcycle.	ORAIS	Vehicles
q_3513	Read a car magazine or book.	ORAIS	Vehicles
q_3514	Rode a motorcycle.	ORAIS	Vehicles
q_3515	Made an entry in a diary or journal.	ORAIS	Writing/Remem.
q_3516	Put pictures in a photo album.	ORAIS	Writing/Remem.
q_3517	Worked on a scrap book.	ORAIS	Writing/Remem.
q_3518	Wrote a handwritten letter.	ORAIS	Writing/Remem.
q_3519	Wrote a postcard.	ORAIS	Writing/Remem.
q_3520	Wrote a thank-you note.	ORAIS	Writing/Remem.
q_5637	Shot a gun.	BAI	BRI2
q_5638	Drank four or more soft drinks a day.	BAI	BRI3
q_5639	Lied about my age.	BAI	BRI4
q_5640	Sang in a car or shower.	BAI	BRI9
q_5642	Spent an hour daydreaming.	BAI	BRI12
q_5643	Consulted a professional nutritionist, dietician, or physician about my diet.	BAI	BRI13
q_5644	Tried to stop using alcohol or other drugs.	BAI	BRI18
q_5645	Ended a romantic relationship.	BAI	BRI22
q_5646	Meditated.	BAI	BRI23
q_5647	Arranged a date for a friend.	BAI	BRI34
q_5648	Finished a large project.	BAI	BRI35
q_5649	Fed a stray dog or cat.	BAI	BRI37
q_5650	Drew pictures, paced, or otherwise fidgeted while on the phone.	BAI	BRI41
q_5651	Consulted a lawyer.	BAI	BRI43

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_5652	Had my back rubbed.	BAI	BRI49
q_5653	Argued with someone.	BAI	BRI56
q_5654	Cut my own hair.	BAI	BRI57
q_5655	Made fun of someone.	BAI	BRI58
q_5656	Ate at an all-you-can-eat buffet.	BAI	BRI59
q_5657	Laughed when no one else was doing so.	BAI	BRI61
q_5658	Did a favor for a friend.	BAI	BRI62
q_5659	Took antacids.	BAI	BRI63
q_5660	Was consulted for help or advice by someone with a personal problem.	BAI	BRI64
q_5661	Thought about work in my free time.	BAI	BRI66
q_5662	Paid someone to polish my shoes.	BAI	BRI67
q_5663	Placed a long distance call to another country.	BAI	BRI69
q_5664	Did something I thought I would never do.	BAI	BRI72
q_5665	Littered.	BAI	BRI73
q_5666	Left a place because of cigarette smoke.	BAI	BRI76
q_5667	Bought a gift for someone.	BAI	BRI80
q_5669	Swore (used offensive language) around other people.	BAI	BRI84
q_5670	Drove a car after having a few alcoholic drinks.	BAI	BRI85
q_5671	Complained about service in a restaurant.	BAI	BRI89
q_5672	Had a beauty treatment or had my hair styled.	BAI	BRI92
q_5673	Attended a religious revival meeting.	BAI	BRI93
q_5674	Changed the place where I live.	BAI	BRI94
q_5675	Hung up the phone on a friend or relative during an argument.	BAI	BRI95
q_5676	Talked in a language other than my native language.	BAI	BRI96
q_5677	Went to a grocery store.	BAI	BRI97
q_5678	Went roller skating, ice skating, or rollerblading.	BAI	BRI102
q_5679	Used a thermometer to take my temperature.	BAI	BRI103
q_5682	Drank alcohol during working hours.	BAI	BRI108
q_5683	Drove while talking on the phone.	BAI	BRI111

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_5684	Had a vaccination shot (such as a flu shot, allergy shot, or tetanus shot).	BAI	BRI112
q_5685	Reported someone to the authorities for some form of misbehavior.	BAI	BRI114
q_5686	Broke a promise.	BAI	BRI115
q_5687	Bought something from a phone or door solicitor.	BAI	BRI117
q_5688	Had my cholesterol level checked.	BAI	BRI119
q_5689	Took a sleeping pill.	BAI	BRI120
q_5690	Shopped at a second-hand or thrift store.	BAI	BRI121
q_5691	Hit or slapped someone.	BAI	BRI123
q_5692	Chauffeured (drove) a child around.	BAI	BRI125
q_5693	Bought a piece of artwork.	BAI	BRI126
q_5694	Took a nap during the day.	BAI	BRI128
q_5695	Rode a bicycle or motorcycle without a helmet.	BAI	BRI130
q_5696	Ate breakfast in bed (not as a patient).	BAI	BRI131
q_5697	Lost my temper.	BAI	BRI134
q_5698	Left a place because it was too crowded.	BAI	BRI136
q_5699	Had an alcoholic drink before breakfast or instead of breakfast.	BAI	BRI137
q_5700	Paid someone to clean house or do yard work.	BAI	BRI138
q_5701	Lounged around my house without any clothes on.	BAI	BRI139
q_5702	Slept past noon.	BAI	BRI141
q_5703	Changed my plans because of weather conditions.	BAI	BRI142
q_5704	Yelled at a stranger.	BAI	BRI143
q_5705	Read about, discussed, or researched sports teams for more than one hour per day.	BAI	BRI144
q_5706	Used smokeless tobacco (such as chewing tobacco or snuff).	BAI	BRI146
q_5707	Donated blood.	BAI	BRI150
q_5708	Talked to a neighbor.	BAI	BRI152
q_5713	Took a long walk alone.	BAI	BRI174
q_5715	Made a new friend.	BAI	BRI179
q_5716	Played in or conducted a band or orchestra.	BAI	BRI180
q_5717	Yelled at an animal.	BAI	BRI184

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_5718	Drank alcohol or used other drugs to make myself feel better.	BAI	BRI186
q_5719	Complimented someone.	BAI	BRI187
q_5720	Slept more than 10 hours at a time.	BAI	BRI188
q_5721	Forgot the birthday of a close friend or relative.	BAI	BRI191
q_5722	Did a physical therapy or rehabilitation session.	BAI	BRI193
q_5723	Changed jobs.	BAI	BRI195
q_5725	Destroyed or damaged an object in anger or frustration.	BAI	BRI198
q_5726	Drove or rode in a car without a seatbelt.	BAI	BRI199
q_5727	Used sunscreen.	BAI	BRI200
q_5728	Told a joke.	BAI	BRI203
q_5729	Eliminated a food from my diet because of health concerns.	BAI	BRI207
q_5730	Rode a bicycle.	BAI	BRI208
q_5731	Started a conversation with strangers.	BAI	BRI210
q_5732	Bought new clothes.	BAI	BRI212
q_5733	Shared a problem with a close friend or relative.	BAI	BRI213
q_5735	Stayed late at work.	BAI	BRI215
q_5736	Made a gift for someone.	BAI	BRI221
q_5737	Visited a psychiatrist or psychologist.	BAI	BRI222
q_5738	Ate until I felt sick.	BAI	BRI223
q_5739	Stayed up all night.	BAI	BRI224
q_5740	Cared for a sick relative.	BAI	BRI225
q_5743	Washed or polished a car (not as a job).	BAI	BRI235
q_5744	Bit my fingernails.	BAI	BRI236
q_5745	Took music lessons (voice or instrument).	BAI	BRI239
q_5746	Borrowed something and lost it, broke it, or never returned it.	BAI	BRI240
q_5748	Apologized to someone.	BAI	BRI249
q_5749	Picked up a hitch-hiker.	BAI	BRI252
q_5750	Took a laxative.	BAI	BRI255
q_5751	Taught Sunday school.	BAI	BRI256
q_5752	Told a dirty joke.	BAI	BRI258
q_5753	Took medication for depression.	BAI	BRI259
q_5755	Yelled at a child.	BAI	BRI262

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_5756	Changed or canceled an appointment.	BAI	BRI264
q_5757	Changed my daily routine because of pain associated with an injury or illness.	BAI	BRI266
q_5758	Discussed sexual matters with a female friend.	BAI	BRI267
q_5759	Read my horoscope.	BAI	BRI274
q_5760	Laughed so hard tears came out of my eyes.	BAI	BRI276
q_5761	Tried to get a tan.	BAI	BRI279
q_5763	Followed a news story closely.	BAI	BRI282
q_5764	Ate meat cooked rare.	BAI	BRI287
q_5765	Tried to convince someone to change his or her religious or political beliefs.	BAI	BRI288
q_5766	Arrived at an event more than an hour late.	BAI	BRI289
q_5767	Visited a person in a hospital.	BAI	BRI291
q_5768	Bought new furniture.	BAI	BRI298
q_5769	Hung up on a phone solicitor.	BAI	BRI299
q_5770	Skipped a meal.	BAI	BRI300
q_5771	Played table tennis or ping-pong.	BAI	BRI301
q_5773	Tried to quit smoking.	BAI	BRI304
q_5774	Ate something spicy for breakfast.	BAI	BRI305
q_5775	Drank black coffee (no cream or sugar).	BAI	BRI307
q_5776	Quit my job.	BAI	BRI309
q_5777	Discussed ways to make money.	BAI	BRI311
q_5778	Smoked a cigarette or cigar before breakfast.	BAI	BRI312
q_5779	Dried flowers or herbs.	BAI	BRI315
q_5780	Played sick to avoid doing something unpleasant.	BAI	BRI317
q_5781	Gestured or honked angrily at the driver of a car.	BAI	BRI318
q_5782	Used eyeglasses or contact lenses.	BAI	BRI319
q_5783	Cheered loudly at a sports event.	BAI	BRI321
q_5784	Gave a tip of more than 20% for some service.	BAI	BRI323
q_5785	Had a professional massage.	BAI	BRI326
q_5786	Let someone else win a game.	BAI	BRI327
q_5787	Rode a horse.	BAI	BRI328
q_5788	Hugged someone.	BAI	BRI329
q_5789	Let work pile up until just before a deadline.	BAI	BRI332

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_5790	Called in sick to work because I was too tired to get up.	BAI	BRI335
q_5791	Repaired or did maintenance on a car myself (not as a job).	BAI	BRI336
q_5792	Decorated a room (not as a job).	BAI	BRI338
q_5793	Discussed sexual matters with a male friend.	BAI	BRI340
q_5796	Played a practical joke on someone.	BAI	BRI344
q_5797	Changed clothes during the work day (excluding gym or athletics).	BAI	BRI346
q_5798	Dieted to lose weight.	BAI	BRI348
q_5799	Drank five or more cups of coffee per day.	BAI	BRI349
q_5800	Laughed out loud at something I thought of.	BAI	BRI350
q_5801	Rode on a roller coaster, Ferris wheel, merry-go-round, or similar ride.	BAI	BRI353
q_5802	Picked up a date in a bar, restaurant, or similar place.	BAI	BRI355
q_5803	Used a sauna or hot tub (whirlpool).	BAI	BRI356
q_5804	Colored my hair.	BAI	BRI360
q_5805	Took vitamins or other health supplements.	BAI	BRI362
q_5807	Made repairs around the house.	BAI	BRI367
q_5808	Chewed on a pen or pencil.	BAI	BRI369
q_5809	Did an imitation or impersonation of another person.	BAI	BRI370
q_5810	Drank tea.	BAI	BRI375
q_5811	Had my blood pressure taken.	BAI	BRI378
q_5812	Sang in or conducted a choir or small ensemble.	BAI	BRI381
q_5813	Took three or more different medications in the same day.	BAI	BRI382
q_5814	Received public assistance (such as food stamps or welfare).	BAI	BRI386
q_5816	Slept less than six hours in a night.	BAI	BRI389
q_5817	Stayed away from a social event in order to finish some work.	BAI	BRI390
q_5818	Drank eight or more glasses of water a day.	BAI	BRI391
q_5819	Carried a good luck charm (such as a rabbit foot or four leaf clover).	BAI	BRI393
q_5820	Drove faster than normal because I was angry.	BAI	BRI394

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_5822	Attended a reunion (school or family).	BAI	BRI398
q_5823	Visited a cemetary or attended a funeral.	BAI	BRI399
q_5824	Attended a fashion show.	BAI	BRI160
q_5825	Went to a movie.	BAI	BRI182
q_5826	Tried on clothes in a store.	BAI	BRI281
q_5827	Painted my nails or toenails.	BAI	BRI5
q_5828	Visited a friend or family member in another city.	BAI	BRI8
q_5829	Spent time preserving, canning, or pickling food.	BAI	BRI16
q_5830	Gave money to a stranger who asked me for money.	BAI	BRI20
q_5831	Was late for work, class, or other responsibilities.	BAI	BRI24
q_5832	Participated in a self-help or support group.	BAI	BRI30
q_5833	Cried nearly every day for a week.	BAI	BRI31
q_5834	Had a medical operation.	BAI	BRI32
q_5835	Took anti-anxiety drugs.	BAI	BRI38
q_5836	Cared for pet fish.	BAI	BRI39
q_5837	Went to a dentist.	BAI	BRI47
q_5838	Had people over to watch a TV show or movie.	BAI	BRI48
q_5839	Did a self-examination for cancer.	BAI	BRI60
q_5840	Borrowed clothing from or lent clothing to a friend.	BAI	BRI65
q_5841	Misplaced something important (such as eyeglasses or car keys).	BAI	BRI71
q_5842	Ate raw fish (such as sushi) or shellfish (such as oysters).	BAI	BRI78
q_5845	Smoked tobacco.	BAI	BRI88
q_5846	Placed a classified ad, Craigslist listing, or similar advertisement.	BAI	BRI90
q_5847	Rode in a taxi or rideshare (such as Lyft or Uber)	BAI	BRI91
q_5848	Was bothered enough by cigarette smoke to take action.	BAI	BRI106
q_5849	Went to a street fair, farmers market, or similar event.	BAI	BRI110

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	Item	Origin	Facet/BAI ID
q_5850	Used OkCupid, Tinder, Grindr, or other dating apps or websites.	BAI	BRI133
q_5851	Took aspirin, ibuprofen, or other mild pain relievers.	BAI	BRI145
q_5852	Visited a doctor for a physical examination or general check up.	BAI	BRI147
q_5853	Did not pay a bill on time.	BAI	BRI148
q_5854	Took a hard drug recreationally (such as cocaine, methamphetamine, or heroin).	BAI	BRI151
q_5856	Went bowling.	BAI	BRI166
q_5857	Had a mammogram or my prostate checked.	BAI	BRI169
q_5858	Asked questions in a meeting, lecture, or other presentation.	BAI	BRI170
q_5859	Did not return a phone call or text from someone I knew.	BAI	BRI177
q_5860	Texted on a cell phone.	BAI	BRI183
q_5861	Took a vacation of one week or more.	BAI	BRI192
q_5862	Talked on the phone for ten minutes or more.	BAI	BRI194
q_5863	Bought organic food or drink.	BAI	BRI201
q_5865	Ate food from a different culture.	BAI	BRI204
q_5866	Drove a car 10 miles (16 km) per hour over the speed limit.	BAI	BRI211
q_5867	Subscribed to a magazine, blog, newspaper, or other print or online media.	BAI	BRI217
q_5868	Ate tuna, halibut, salmon, or another fish.	BAI	BRI229
q_5869	Had a Pap smear.	BAI	BRI251
q_5870	Had an eye examination.	BAI	BRI253
q_5871	Attended a city council, student council, or other similar administrative meeting.	BAI	BRI270
q_5872	Drove more than 200 miles (325 km) by myself.	BAI	BRI271
q_5873	Went to a garage sale, yard sale, rummage sale or similar event.	BAI	BRI272
q_5874	Borrowed a substantial amount of money from a friend or family member.	BAI	BRI277
q_5875	Made a list to help myself organize.	BAI	BRI278
q_5876	Lent a substantial amount of money to a friend or family member.	BAI	BRI285
q_5877	Skipped work or classes.	BAI	BRI316

Continued on next page

Table B.1 – Items of the BARE Inventory, *continued from previous page*

SAPA ID	**Item**	**Origin**	**Facet/BAI ID**
q_5878	Smoked, vaped or otherwise consumed marijuana.	BAI	BRI324
q_5879	Took cough syrup or cough drops for a cough.	BAI	BRI325
q_5880	Had an overdue fine for a library book or other rental.	BAI	BRI347
q_5882	Went on a blind date.	BAI	BRI368
q_5883	Read a tabloid paper or online tabloid.	BAI	BRI373
q_5884	Paid close attention to my finances.	BAI	BRI377
q_5885	Learned how to use a new computer program.	BAI	BRI28
q_5887	Worked crossword puzzles, Sudoku, or similar puzzles.	BAI	BRI118
q_5889	Drank an energy drink or took caffeine pills.	BAI	BRI293
q_5890	Read the funny pages or comics (paper or online).	BAI	BRI380
q_5891	Attended a wedding.	BAI	BRI294

Table B.2: Personality items removed from BISCUIT's analysis of best items for a criterion due to the item content being synonymous with the criterion. Item correlation with the corresponding criterion is listed.

Item	Item Pool	Criterion	Corr.
Feel healthy and vibrant most of the time.	Traditional	Health	.51
Get stressed out easily.	Traditional	Stress	.43
Smoked tobacco.	Behavioral	Smoking	.88
Smoked a cigarette or cigar before breakfast.	Behavioral	Smoking	.85
Tried to quit smoking.	Behavioral	Smoking	.52
Exercised for 40 minutes or longer.	Behavioral	Exercise	.69
Did aerobic exercise.	Behavioral	Exercise	.58
Participated in an exercise program.	Behavioral	Exercise	.57
Used an exercise machine.	Behavioral	Exercise	.48
Lifted weights.	Behavioral	Exercise	.48
Did yoga or other movement exercises.	Behavioral	Exercise	.47
Went running or jogging.	Behavioral	Exercise	.40

Table B.3: Reliabilities of BISCUIT models as if they were personality scales, by criterion. Items used per model and correlation with appropriate criterion are also listed.

Criterion	Item Pool	Items Used	Corr.	α	ω_h	ω_{total}
Health	Traditional	10	.48	.89	.64	.93
	Behaviors	10	.46	.76	.77	.82
	Combined	10	.51	.86	.63	.91
Stress	Traditional	20	.52	.94	.71	.96
	Behaviors	10	.35	.71	.48	.78
	Combined	20	.52	.94	.71	.96
BMI	Traditional	41	.42	.77	.57	.83
	Behaviors	10	.44	.72	.48	.78
	Combined	14	.48	.70	.50	.79
Smoking	Traditional	26	.29	.89	.64	.91
	Behaviors	10	.54	.79	.62	.85
	Combined	10	.53	.81	.67	.87
Exercise	Traditional	27	.41	.94	.60	.95
	Behaviors	10	.49	.73	.57	.77
	Combined	10	.51	.80	.46	.88
ER Visits	Traditional	10	.19	.79	.48	.87
	Behaviors	10	.24	.72	.49	.79
	Combined	10	.29	.51	.40	.70

Table B.4: The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with general health ($R = .48$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Tire out quickly.	−.39	IPIP-HEXACO	Ext./Liveliness
Have great stamina.	.38	IPIP-HEXACO	Ext./Liveliness
Am usually active and full of energy.	.36	IPIP-HEXACO	Ext./Liveliness
Often feel listless and tired for no reason.	−.34	EPQ	Neuroticism
Recover quickly from stress and illness.	.34	QB6	Resiliency
Am happy with my life.	.34	QB6	Resiliency
Feel a sense of worthlessness or hopelessness.	−.32	QB6	Resiliency
Have a low opinion of myself.	−.32	IPIP-NEO	Neur./Depression
Feel that I'm unable to deal with things.	−.32	IPIP-NEO	Neur./Vulnerability
Dislike myself.	−.31	IPIP-NEO	Neur./Depression

*Ext. = Extraversion; Neur. = Neuroticism

Table B.5: The 10 personality items most strongly correlated with **general health**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with general health ($R = .46$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Did aerobic exercise.	.31	BARE (ORAIS)	Exercise
Exercised for 40 minutes or longer.	.31	BARE (ORAIS)	Exercise
Participated in an exercise program.	.31	BARE (ORAIS)	Exercise
Lifted weights.	.26	BARE (ORAIS)	Exercise
Went running or jogging.	.26	BARE (ORAIS)	Exercise
Changed my daily routine because of pain associated with an injury or illness.	−.25	BARE	None
Went on a hike.	.24	BARE (ORAIS)	Summer Activities
Used an exercise machine.	.23	BARE (ORAIS)	Exercise
Cried nearly every day for a week.	−.23	BARE	None
Took three or more different medications in the same day.	−.23	BARE	None

Table B.6: The 20 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with overall stress ($R = .52$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Find life difficult.	.41	Plasticity/Stability	Stability
Am relaxed most of the time.	−.40	IPIP-NEO	Neur./Anxiety
Feel a sense of worthlessness or hopelessness.	.37	QB6	Resiliency
Feel desperate.	.37	IPIP-NEO	Neur./Depression
Recover quickly from stress and illness.	−.37	QB6	Resiliency
Am happy with my life.	−.37	QB6	Resiliency
Am often down in the dumps.	.36	IPIP-NEO	Neur./Depression
Often feel blue.	.36	IPIP-NEO	Neur./Depression
Feel healthy and vibrant most of the time.	−.36	IPIP-HEXACO	Ext./Liveliness
Get caught up in my problems.	.36	IPIP-NEO	Neur./Anxiety
Worry about things.	.36	IPIP-NEO	Neur./Anxiety
Often feel fed-up.	.35	EPQ	Neuroticism
Rarely feel depressed.	−.35	BFAS	Neur./Withdrawal
Am often in a bad mood.	.35	IPIP-NEO	Neur./Anger
Dislike myself.	.35	IPIP-NEO	Neur./Depression
Often feel lonely.	.35	EPQ	Neuroticism
Rarely worry.	−.35	IPIP-HEXACO	EmS./Anxiety
Feel that I'm unable to deal with things.	.34	IPIP-NEO	Neur./Vulnerability
Suffer from nerves.	.34	EPQ	Neuroticism
Am a worrier.	.33	EPQ	Neuroticism

*EmS = Emotional Stability; Ext. = Extraversion; Neur. = Neuroticism

Table B.7: The 10 personality items most strongly correlated with **overall stress**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a moderate correlation with overall stress ($R = .35$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Cried nearly every day for a week.	.28	BARE	None
Lost my temper.	.23	BARE	None
Destroyed or damaged an object in anger or frustration.	.19	BARE	None
Let work pile up until just before a deadline.	.19	BARE	None
Played sick to avoid doing something unpleasant.	.18	BARE	None
Visited a psychiatrist or psychologist.	.17	BARE	None
Did not return a phone call or text from someone I knew.	.16	BARE	None
Was late for work, class, or other responsibilities.	.16	BARE	None
Swore (used offensive language) around other people.	.16	BARE	None
Argued with someone.	.15	BARE	None

Table B.8: The 41 personality items most strongly correlated with **BMI**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with BMI ($R = .42$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Often eat too much.	.34	IPIP-NEO	Neur./Immoderation
Am able to control my cravings.	−.24	IPIP-NEO	Neur./Immoderation
Rarely overindulge.	−.20	IPIP-NEO	Neur./Immoderation
Don't strive for elegance in my appearance.	.17	IPIP-HEXACO	Hon./Greed avoidance
Love to eat.	.17	IPIP-NEO	Neur./Immoderation
Am usually active and full of energy.	−.17	IPIP-HEXACO	Ext./Liveliness
Never spend more than I can afford.	−.16	IPIP-NEO	Neur./Immoderation
Feel little concern for others.	−.15	IPIP-BFFM	Agreeableness
Would never go hang-gliding or bungee-jumping.	.15	IPIP-NEO	Ext./Excitement seeking
Have great stamina.	−.14	IPIP-HEXACO	Ext./Liveliness
Never splurge.	−.14	IPIP-NEO	Neur./Immoderation
Believe one has special duties to one's family.	.14	EPQ	Psychoticism
Seek adventure.	−.14	IPIP-NEO	Ext./Excitement seeking
Easily resist temptations.	−.14	IPIP-NEO	Neur./Immoderation
Feel healthy and vibrant most of the time.	−.13	IPIP-HEXACO	Ext./Liveliness
Can't make up my mind.	−.13	IPIP-NEO	Neur./Vulnerability
Am a person whose moods go up and down easily.	−.12	BFAS	Neur./Volatility
Do the opposite of what is asked.	−.12	IPIP-NEO	Con./Dutifulness
Need protection.	−.12	IPIP-HEXACO	EmS./Dependence
Love excitement.	−.12	IPIP-NEO	Ext./Excitement seeking
Like to arrive at appointments in plenty of time.	.12	EPQ	Psychoticism
Seek status.	−.12	IPIP-HEXACO	Hon./Greed avoidance
Tremble in dangerous situations.	−.11	IPIP-HEXACO	EmS./Fearfulness
Admire a really clever scam.	−.11	IPIP-HEXACO	Hon./Fairness
Am seldom bothered by the apparent suffering of strangers.	−.11	IPIP-HEXACO	EmS./Sentimentality
Am able to stand up for myself.	.11	IPIP-NEO	Neur./Self-consciousness
Like to do frightening things.	−.11	IPIP-HEXACO	EmS./Fearfulness
Make careless mistakes.	−.11	IPIP-HEXACO	Con./Prudence
Prefer to go my own way rather than act by the rules.	−.11	EPQ	Psychoticism
Am a creature of habit.	.10	IPIP-NEO	Open./Adventurousness
Get even with others.	−.10	IPIP-HEXACO	Agr./Forgiveness
Love dangerous situations.	−.10	IPIP-HEXACO	EmS./Fearfulness
Enjoy being reckless.	−.10	IPIP-NEO	Ext./Excitement seeking
Seek danger.	−.10	IPIP-NEO	Ext./Excitement seeking
Go on binges.	.10	IPIP-NEO	Neur./Immoderation
Don't understand things.	−.10	IPIP-NEO	Con./Self-efficacy
Find fault with everything.	−.10	IPIP-HEXACO	Agr./Gentleness
Love surprise parties.	−.10	IPIP-NEO	Ext./Gregariousness
Like to attract attention.	−.10	IPIP-HEXACO	Hon./Modesty
Need reassurance.	−.10	IPIP-HEXACO	EmS./Dependence
Would call myself a nervous person.	−.10	EPQ	Neuroticism

*Agr. = Agreeableness; Con. = Conscientiousness; EmS = Emotional Stability; Ext. = Extraversion; Hon. = Honesty-Humility; Neur. = Neuroticism; Open. = Openness

Table B.9: The 10 personality items most strongly correlated with **BMI**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with BMI ($R = .44$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Ate too much.	.26	BARE (ORAIS)	Food-Related
Dieted to lose weight.	.26	BARE	None
Had my cholesterol level checked.	.24	BARE	None
Used public transportation.	−.24	BARE (ORAIS)	Green Activities
Consulted a professional nutritionist, dietician, or physician about my diet.	.24	BARE	None
Ate or drank while driving.	.23	BARE (ORAIS)	Food-Related
Took antacids.	.23	BARE	None
Took three or more different medications in the same day.	.22	BARE	None
Had my blood pressure taken.	.21	BARE	None
Bought a car, truck, or motorcycle.	.21	BARE (ORAIS)	Vehicles

Table B.10: The 26 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a moderate correlation with smoking frequency ($R = .29$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Would take drugs which may have strange or dangerous effects.	.32	EPQ	Psychoticism
Have some bad habits.	.20	IPIP-MPQ	Unlikely virtues
Go on binges.	.17	IPIP-NEO	Neur./Immoderation
Have often gone against my parents wishes.	.17	EPQ	Psychoticism
People have said that I sometimes act rashly.	.17	EPQ	Extraversion
Am self-destructive.	.16	Plasticity/Stability	Stability
Never spend more than I can afford.	−.15	IPIP-NEO	Neur./Immoderation
Cheat on people who have trusted me.	.15	IPIP-HEXACO	Hon./Fairness
Act wild and crazy.	.15	IPIP-NEO	Ext./Excitement seeking
Do dangerous things.	.15	IPIP-MPQ	Harm avoidance
Try to follow the rules.	−.15	IPIP-NEO	Con./Dutifulness
Enjoy being reckless.	.15	IPIP-NEO	Ext./Excitement seeking
Take risks that could cause trouble for me.	.14	QB6	Honesty-Propriety
Rebel against authority.	.14	IPIP-HEXACO	Open./Unconventionality
Make rash decisions.	.14	IPIP-NEO	Con./Cautiousness
Often feel listless and tired for no reason.	.13	EPQ	Neuroticism
Avoid dangerous situations.	−.13	IPIP-MPQ	Harm avoidance
Feel healthy and vibrant most of the time.	−.13	IPIP-HEXACO	Ext./Liveliness
Break rules.	.13	IPIP-NEO	Con./Dutifulness
Dislike loud music.	−.13	IPIP-NEO	Ext./Excitement seeking
Can easily get some life into a dull party.	.13	EPQ	Extraversion
Do things that I later regret.	.13	IPIP-NEO	Neur./Immoderation
Do crazy things.	.13	IPIP-NEO	Con./Cautiousness
Like to be viewed as proper and conventional.	−.13	IPIP-HEXACO	Open./Unconventionality
Pay my bills on time.	−.13	IPIP-NEO	Con./Dutifulness
Respect authority.	−.13	BFAS	Agr./Politeness

*Agr. = Agreeableness; Con. = Conscientiousness; Ext. = Extraversion; Hon. = Honesty-Humility; Neur. = Neuroticism; Open. = Openness

Table B.11: The 10 personality items most strongly correlated with **smoking frequency**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with smoking frequency ($R = .54$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Smoked, vaped or otherwise consumed marijuana.	.50	BARE	None
Drank alcohol or used other drugs to make myself feel better.	.37	BARE	None
Took a hard drug recreationally (such as cocaine, methamphetamine, or heroin).	.33	BARE	None
Had a hangover.	.32	BARE (ORAIS)	Drinking
Left a place because of cigarette smoke.	−.32	BARE	None
Became intoxicated.	.28	BARE (ORAIS)	Drinking
Tried to stop using alcohol or other drugs.	.27	BARE	None
Used smokeless tobacco (such as chewing tobacco or snuff).	.27	BARE	None
Had an alcoholic drink before breakfast or instead of breakfast.	.25	BARE	None
Drank five or more cups of coffee per day.	.24	BARE	None

Table B.12: The 27 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .41$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Feel healthy and vibrant most of the time.	.35	IPIP-HEXACO	Ext./Liveliness
Am usually active and full of energy.	.33	IPIP-HEXACO	Ext./Liveliness
Tire out quickly.	−.32	IPIP-HEXACO	Ext./Liveliness
Have great stamina.	.31	IPIP-HEXACO	Ext./Liveliness
Often feel listless and tired for no reason.	−.30	EPQ	Neuroticism
Do a lot in my spare time.	.29	IPIP-NEO	Ext./Activity level
Am easily discouraged.	−.27	BFAS	Neur./Withdrawal
Am afraid of many things.	−.26	IPIP-NEO	Neur./Anxiety
Maintain high energy throughout the day.	.25	IPIP-HEXACO	Ext./Liveliness
Feel that my life lacks direction.	−.25	IPIP-NEO	Neur./Depression
Sometimes feel just miserable for no reason.	−.25	EPQ	Neuroticism
Feel a sense of worthlessness or hopelessness.	−.24	QB6	Resiliency
Hang around doing nothing.	−.23	IPIP-HEXACO	Con./Diligence
Have little to contribute.	−.23	IPIP-NEO	Con./Self-efficacy
Turn plans into actions.	.23	IPIP-NEO	Con./Achievement striving
Feel short-changed in life.	−.23	IPIP-MPQ	Alienation
Need a push to get started.	−.23	IPIP-NEO	Con./Self-discipline
Do things in a half-way manner.	−.23	IPIP-BFFM	Conscientiousness
Often feel life is very dull.	−.23	EPQ	Neuroticism
Habitually blow my chances.	−.22	Plasticity/Stability	Stability
Am good at many things.	.22	IPIP-BFFM	Intellect
Feel that I'm unable to deal with things.	−.22	IPIP-NEO	Neur./Vulnerability
Break my promises.	−.22	IPIP-NEO	Con./Dutifulness
Waste my time.	−.22	IPIP-NEO	Con./Self-discipline
Am happy with my life.	.22	QB6	Resiliency
Do just enough work to get by.	−.22	IPIP-NEO	Con./Achievement striving
Neglect my duties.	−.21	IPIP-BFFM	Conscientiousness

*Con. = Conscientiousness; Ext. = Extraversion; Neur. = Neuroticism

Table B.13: The 10 personality items most strongly correlated with **exercise frequency**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a large correlation with exercise frequency ($R = .49$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Went on a hike.	.37	BARE (ORAIS)	Summer Activities
Took a long walk alone.	.30	BARE	None
Attended an athletic event.	.26	BARE (ORAIS)	Sports
Recycled one or more items.	.26	BARE (ORAIS)	Green Activities
Bought organic food or drink.	.25	BARE	None
Used sunscreen.	.25	BARE	None
Slept past noon.	−.25	BARE	None
Went swimming.	.24	BARE (ORAIS)	Summer Activities
Did yard work.	.23	BARE (ORAIS)	Gardening
Went sightseeing.	.23	BARE (ORAIS)	Travel

Table B.14: The 10 personality items most strongly correlated with **ER visits**, selected by BISCUIT from a pool of **696 traditional personality items.** The BISCUIT model composed of these items had a moderate correlation with ER visits ($R = .19$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet*
Don't know why I do some of the things I do.	.13	IPIP-NEO	Neur./Immoderation
Find myself in the same kinds of trouble, time after time.	.13	Plasticity/Stability	Stability
Am self-destructive.	.12	Plasticity/Stability	Stability
Have great stamina.	−.11	IPIP-HEXACO	Ext./Liveliness
Rush into things.	.11	IPIP-NEO	Con./Cautiousness
Recover quickly from stress and illness.	−.11	QB6	Resiliency
Feel desperate.	.10	IPIP-NEO	Neur./Depression
Get stressed out easily.	.10	IPIP-NEO	Neur./Anxiety
Feel short-changed in life.	.10	IPIP-MPQ	Alienation
Suffer from sleeplessness.	.10	EPQ	Neuroticism

*Con. = Conscientiousness; Ext. = Extraversion; Neur. = Neuroticism

Table B.15: The 10 personality items most strongly correlated with **ER visits**, selected by BISCUIT from a pool of **425 behavioral frequencies.** The BISCUIT model composed of these items had a moderate correlation with ER visits ($R = .24$). Each listed correlation is an average across ten folds, and its standard error $\approx .02$.

Item	Corr.	Inventory	Domain/Facet
Had my blood pressure taken.	.17	BARE	None
Took three or more different medications in the same day.	.16	BARE	None
Visited a doctor for a physical examination or general check up.	.16	BARE	None
Cried nearly every day for a week.	.13	BARE	None
Changed my daily routine because of pain associated with an injury or illness.	.13	BARE	None
Had a medical operation.	.12	BARE	None
Used a thermometer to take my temperature.	.12	BARE	None
Took anti-anxiety drugs.	.11	BARE	None
Took a sleeping pill.	.10	BARE	None
Took medication for depression.	.10	BARE	None