Laying personality BARE: Behavioral frequencies strengthen personality-criterion relationships

Lorien G. Elleman[1], David M. Condon[2], and William Revelle[1]

1. Department of Psychology, Northwestern University, Evanston, Illinois. 2. Department of Psychology, University of Oregon, Eugene, Oregon.

Abstract

Personality consists of stable patterns of cognitions, emotions, and behaviors, yet personality psychologists rarely study behaviors. Even when examined, behaviors typically are considered to be validation criteria for traditional personality items. In the current study ($N = 332,489$), we conceptualize (self-reported, yearlong) behavioral frequencies as measures of personality. We investigate whether behavioral frequencies have incremental validity over 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 combined pool are more strongly correlated to four criteria. We find mixed evidence of 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 frequencies). Findings suggest that behavioral frequencies are measures of personality that offer a unique effect in describing personality-criterion relationships beyond traditional personality items. We provide an updated, public-domain item pool of behavioral frequencies: the BARE (Behavioral Acts, Revised and Expanded) Inventory.

Keywords

Behavioral frequency; statistical learning; machine learning; nuance; behavior

Introduction

Of the three patterns commonly associated with personality (cognitions, emotions, and behaviors),[1] behaviors are the least studied (Furr, 2009). A criticism of psychology is that it "mostly involves asking people to report on their thoughts, feelings, memories, and attitudes" (Baumeister, Vohs, & Funder, 2007, p. 396) by filling out personality inventories, such as the BFI-2 (Soto & John, 2017), the IPIP-HEXACO (Ashton, Lee, & Goldberg, 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 (usually themselves), using a Likert-like scale (e.g., "Strongly disagree" to "Strongly agree").

Some traditional personality items attempt to measure behaviors; Wilt (2014) found that, of the Big Five domains, judges tended to rate Conscientiousness and Extraversion items as having the most behavioral content (e.g., "Get chores done right away" and "Talk to a lot of different people at parties," respectively). These types of behavioral items lack precision in two key ways. First, the frequency of each behavior is ill-defined, leaving the participant to decide for themselves how their agreement or disagreement with a personality item corresponds to how often they have behaved a certain way in the real world. Second, no time period is specified for a given behavioral pattern; again, it is up to the participant to decide how to connect the personality item to the target's actual behavior. Both of these problems introduce cognitions and emotions into the measurement of behaviors; instead of only reporting their behaviors, participants also report their thoughts and feelings about their behaviors.

The Act Frequency Approach (AFA; Buss & 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 accurate measurements of behavior because they quantify the frequency of each behavior and specify a time period for each frequency. Additionally, behavioral frequencies may represent a more comprehensive range of

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

[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.

behaviors; unlike traditional personality items, behavioral frequencies are not limited to a subset of behaviors that fit neatly into the factor structure of broader personality traits. Despite the potential advantages that behavioral frequencies may have over traditional personality items, they have rarely been studied since the theoretical foundation of AFA (namely, that there is no explanatory power in personality traits because they are merely behavioral summaries) was met with criticism in the late 1980s (e.g., Block, 1989; Moser, 1989; for a review of AFA theory, see Buss & Craik, 1983). By abandoning the behavioral frequency because of AFA's perhaps flawed theory, personality psychology threw out the baby with the bath water; although there is a historical association between the two, the study of the behavioral frequency does not require the baggage of AFA theory.

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 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, John, Craik, & Robins, 1998; Vazire & Mehl, 2008; Jackson et al., 2010).

Post-AFA researchers have not considered behavioral frequencies to be measures of personality, but instead validation criteria for traditional personality traits. For example: Grucza & Goldberg (2007) selected behavioral frequencies as one of several criteria to test the comparative validity of eleven personality inventories; Hirsh, DeYoung, & Peterson (2009) found behavioral patterns for two metatraits (higher-order traits that supposedly subsume the Big Five); Church, Katigbak, Miramontes, del Prado, & Cabrera (2007) found cross-cultural consistency of associations between

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

behavioral frequencies and Big Five traits; Chapman & Goldberg (2017) described behavioral "signatures" of each Big Five trait; and Skimina, Cieciuch, Schwartz, Davidov, & Algesheimer (2019) linked a person's values with their behavior. Since the AFA, however, no published paper has considered that behavioral frequencies measure personality and may account for variance in real-world criteria that is unexplained by traditional personality items.

Analyzing individual personality items with statistical learning techniques

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, Kandler, Bleidorn, Riemann, & McCrae, 2017; Mõttus et al., 2019) and there is cross-rater agreement concerning their specific variance (Mõttus, McCrae, Allik, & Realo, 2014). Given the same pool of items, item-based models better predict outcomes than models of multi-item scales, because the variance associated with individual items has predictive validity (Seeboth & Mõttus, 2018; Mõttus et al., 2015; Mõttus, Bates, Condon, Mroczek, & Revelle, 2017; Revelle, Dworak, & Condon, 2020).

Compared to the study of broad traits, item-level analyses require a large number of multiple comparisons, which, without statistical adjustment, can lead to models that overfit the sample and/or contain rampant false positives. Statistical learning techniques (SLTs) are methods that are more complex than traditional statistical analyses used in personality psychology, such as correlation and regression. One goal of this added complexity is to protect against overfitting (for overviews of SLTs in psychology, see Chapman, Weiss, & Duberstein, 2016 and Yarkoni & Westfall, 2017). Recent research has indicated that applying SLTs to item-level analysis is effective in increasing the predictive accuracy of personality data beyond traditional methods (e.g., Seeboth & Mõttus, 2018; Elleman, Condon, McDougald, & Revelle, 2020).

Two pilot studies have examined the incremental validity of behavioral frequencies

Two unpublished pilot studies have compared the predictive validity of yearlong behavioral frequencies over traditional personality items. The first study ($N = 31,467$; Elleman, Condon, &

Revelle, 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, Condon, & Revelle, 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 be-

havior 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.

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., 2020), 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 1 in the supplemental materials).

Methods

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."

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, Wilt, & Rosenthal, 2010; Revelle et al., 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 & Revelle, 2015; Condon, Roney, & Revelle, 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, Saucier, & Eigenhuis, 2011); BFAS (DeYoung, Quilty, & Peterson, 2007); EPQ (Eysenck, Eysenck, & Barrett, 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 & 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 were included in this study than the second pilot study due to a data sharing agreement with the administrator of the SAPA website.

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

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

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).

As is typical with SLT models, it was necessary to set BISCUIT's tuning parameter. The purpose of a tuning parameter (also known as a hyperparameter) is to set boundaries for an SLT's optimization process by giving it a finite range of possible values for a model parameter. The SLT creates a new model for each new allowable value of the tuning parameter, and then finds the optimal model across the range of values of the tuning parameter. BISCUIT's tuning parameter was the number of items allowed in a given model. We set this range 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.

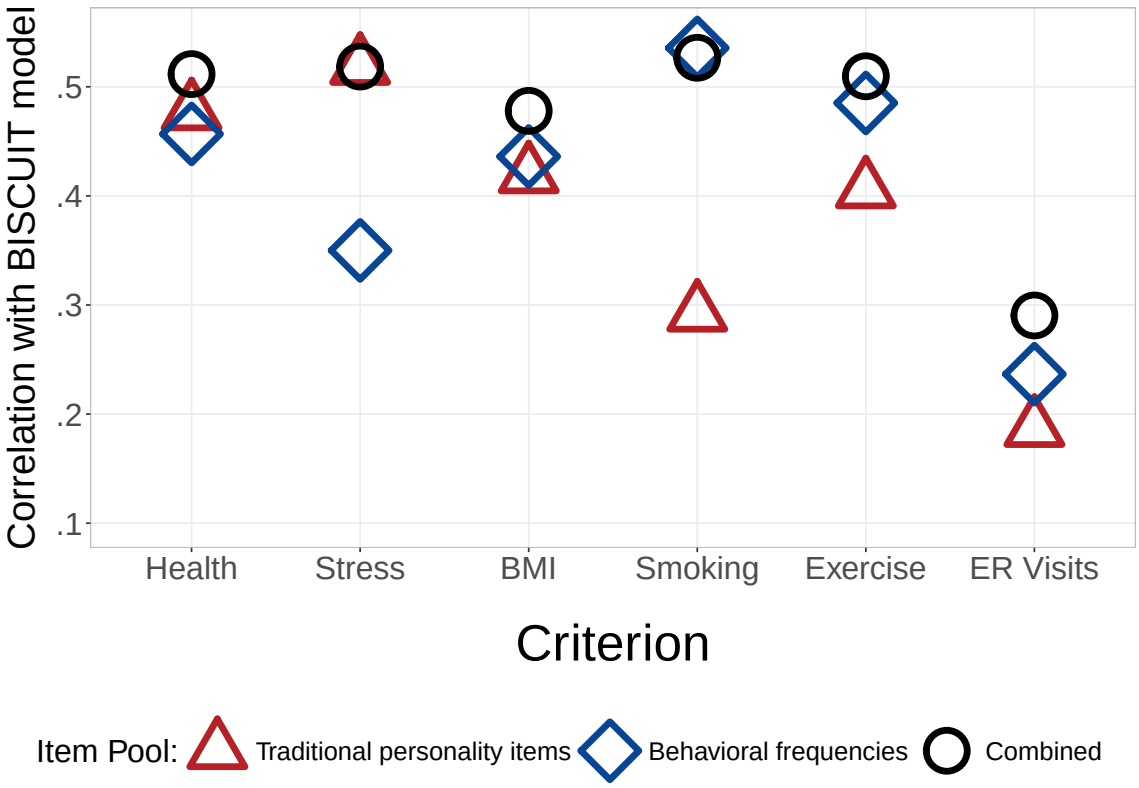
After a BISCUIT model was built and cross-validated on a criterion, we examined the item content of the best items. Any of the best items that were synonymous with the criterion were removed from the pool of possible items and the model was rerun. For example, the behavioral frequency "smoked tobacco" 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 1 in the supplemental materials. For the reliabilities of each BISCUIT model as if it were a typical personality scale, see Table 3 in the supplemental materials.

Results

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 determined whether pairs of BISCUIT models, each of which used a different item pool, were differentially correlated with a criterion, taking into account the fact that there was a non-zero correlation between each pair of models (Steiger, 1980). 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

Figure 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.



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 1; Table 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 1; Table 1).

Table 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**

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 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

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

predominantly associated with traditional personality items; and BMI and smoking frequency with behavioral frequencies (Table 2).

Table 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	–	–

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). Three traditional personality items were from the Liveliness facet of the Extraversion domain (e.g., "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 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").

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

Table 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 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

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

Table 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 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

Table 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	

were seven traditional personality items from four inventories (Table 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 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 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 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 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	−

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.

Three criteria (general health, exercise frequency, and ER visits) were not predominantly asso-

ciated with one type of personality item. Thus, some criteria may be mostly related to a personality item type that is congruent with its type, some criteria may show no such pattern. Interestingly, in this study there were no examples 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 & 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.

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 & 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., 2020). A more complete data structure, or another SLT, such as the elastic net (Zou & 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.

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, Pennebaker, Crow, Dabbs, & Price, 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, Dehghani, & Sagi, 2014;

Chen & 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, Ko, Mannarelli, & Morris, 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 & 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.

Conclusions

Personality is 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.

Appendix

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 & 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 & Goldberg, 2007; Goldberg & Saucier, 2016). Grucza & 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, Visser, Ashton, Lee, & Goldberg, 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 & 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 & 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.[7] 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 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 & 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,[8] and 226 of which originated from the BAI.

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

[8] 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.

References

Ashton, M. C., Lee, K., & 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

Bashaw, W. L., & 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., & 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

Buss, D. M., & 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., & 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., & 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., & 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

Chapman, B. P., & 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., & 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., & 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., & 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. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. doi:10.31234/osf.io/sc4p9

Condon, D. M., & 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., & 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

Csikszentmihalyi, M., & Larson, R. (1987, September). 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

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

DeYoung, C. G., Quilty, L. C., & 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., McDougald, S. K., & 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., & Revelle, W. (2017). Behaviors predict outcomes better than the Big Five. In *The biennial meeting of the association for research in personality.* Sacramento, CA.

Elleman, L. G., Condon, D. M., & Revelle, W. (2018, July). Behavioral frequencies strengthen the link between personality and health behaviors. In *the biennial meeting of the european association for personality psychology.* Zadar, Croatia.

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

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

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

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* (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., & 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., & Saucier, G. (2016). *The Eugene-Springfield Community Sample: Information Available from the Research Participants* (Tech. Rep.). Retrieved from https://ipip.ori.org/ESCS_TechnicalReport_January2016.pdf

Gosling, S. D., John, O. P., Craik, K. H., & 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., & 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

Grucza, R. A., & 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., & Gosling, S. D. (2016, November). Using Smartphones to Collect Behavioral Data in Psychological Science. *Perspectives on Psychological Science*, *11*(6), 838–854. doi:10.1177/1745691616650285

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

Hirsh, J. B., DeYoung, C. G., & 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

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

Hope, A. C. A. (1968, September). 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

Iliev, R., Dehghani, M., & Sagi, E. (2014, July). 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., & 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

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

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

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & 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

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., Bates, T. C., Condon, D. M., Mroczek, D. K., & 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., & 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.supp

Mõttus, R., McCrae, R. R., Allik, J., & 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., Metspalu, A., & 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., ... 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

Pozzebon, J. A., Visser, B. A., Ashton, M. C., Lee, K., & 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

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

Revelle, W. (2020). psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois. Retrieved from `https://CRAN.R-project.org/package=psych`

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

Revelle, W., Dworak, E. M., & 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., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In *Handbook of individual differences in cognition* (pp. 27–49). New York, NY: Springer New York. doi:10.1007/978-1-4419-1210-7_2

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

Seeboth, A., & 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., & 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., & 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

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., ... 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

Thalmayer, A. G., Saucier, G., & 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

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

Wilt, J. (2014). *A New Form and Function for Personality* (Doctoral dissertation). doi:10.1037/e571452013-180

Wilt, J., & 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., & 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., & 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