European Journal of Personality, *Eur. J. Pers.* (2020)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2235

Comparing predictive validity in a community sample: High-dimensionality and traditional domain-and-facet structures of personality variation

GERARD SAUCIER[1]*, KATHRYN IURINO[1] and AMBER GAYLE THALMAYER[2]

[1] *University of Oregon, USA*
[2] *University of Lausanne, Switzerland*

Abstract: Prediction of outcomes is an important way of distinguishing, among personality models, the best from the rest. Prominent previous models have tended to emphasize multiple internally consistent "facet" scales subordinate to a few broad domains. But such an organization of measurement may not be optimal for prediction. Here, we compare the predictive capacity and efficiency of assessments across two types of personality-structure model: conventional structures of facets as found in multiple platforms, and new high-dimensionality structures emphasizing those based on natural-language adjectives, in particular lexicon-based structures of 20, 23, and 28 dimensions. Predictions targeted 12 criterion variables related to health and psychopathology, in a sizeable American community sample. Results tended to favor personality-assessment platforms with (at least) a dozen or two well-selected variables having minimal intercorrelations, without sculpting of these to make them function as indicators of a few broad domains. Unsurprisingly, shorter scales, especially when derived from factor analyses of the personality lexicon, were shown to take a more efficient route to given levels of predictive capacity. Popular 20th-century personality-assessment models set out influential but suboptimal templates, including one that first identifies domains and then facets, which compromise the efficiency of measurement models, at least from a comparative-prediction standpoint. © 2020 European Association of Personality Psychology

Key words: Lexical studies; Personality scales and inventories; Regression methods

One important goal of personality research is to improve methods for personality assessment. This must involve defining well the structure of personality variation, but also refining the techniques by which such variation is translated into meaningful and interpretable scores. Such pursuits will be well-served by comparisons of numerous structures and variations in technique. This report focuses on comparisons of just this sort, aiming to delineate not only which instruments predict better or worse, but also which attributes of instruments are associated with better or worse prediction.

As is well-known, there are a variety of models of personality structure available in multiscale inventories. These vary not only in which traits they measure, but oalso features like how many items, how many scales (or variables), how many items per variable, and how succinct versus wordy the items are. Here, we not only compare inventories globally, but also search in a more granular fashion for the attributes of inventories that might contribute most or least to predictive capacity.

For investigations that further these such searches and comparisons, a desirable data-set would have many personality-assessment models represented with respect to a common set of research participants, and enough relevant criterion variables to enable some useful comparisons. At this time, perhaps the best-suited data for this purpose is found in the Eugene-Springfield Community Sample (ESCS). Between 1993 and 2006, participants were mailed 29 distinct questionnaire-packages, some of them containing a single long questionnaire, others a combination of instruments. At various times during this time-period, a wide variety of personality adjectives were administered as part of various questionnaire-packages, Moreover, also administered during this time were several thousand items that have come to be known as the International Personality Item Pool IPIP; Goldberg et al, (2006), a resource that enables the development of new instruments. Besides personality questionnaires, the sample completed multiple instruments related to health and psychopathology, which affords relevant criterion variables in those important outcome domains.

The ESCS data-set already provided a rich resource for a comparative-prediction study, but we saw an opportunity to enrich this resource and increase informativeness by refining additional models of two different sorts. Thus, before

*Correspondence to: Gerard Saucier, University of Oregon, USA.
E-mail: gsaucier@uoregon.edu

The authors are grateful to Lewis R. Goldberg for years of dedicated work creating and maintaining the Eugene-Springfield Sample and the International Personality Item Pool, and to Carrie Bettenhausen for research assistance. Pregistrations associated with this article are found at: https://osf.io/7an35/, https://osf.io/2dhqb/, https://osf.io/vq8mu/, and https://osf.io/hxtb8/. Eugene-Springfield Community Sample data is accessible via Harvard Dataverse.

This article earned Preregistered + Analysis Plan badge through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The permanent path to the registration is openly accessible at https://osf.io/hxtb8/?view_only=9fa10ba7dc7e47afa2f5b54419e686df. Author's disclosure form may also be found at the Supporting Information in the online version.

Handling editor: Dr EJP Guest Editor
Received 16 April 2019
Revised 10 January 2020, Accepted 13 January 2020

embarking on the main comparative-prediction study, we engaged in two distinct preliminary studies.

The Eugene-Springfield Community Sample has been a primary source for lexicon-based structures of personality. The large collection of adjectives enables representation of models identified in other lexicon-based studies. More focally, in 1995 the sample was administered the 500-PDA (500 person-descriptive adjectives), which preliminary work had indicated to be the 500 most familiar, frequently used attribute-descriptive adjectives in American English. Saucier (1997) drew on 500-PDA data to examine the effects of variable-selection on structures of relatively few (i.e., one to seven) factors, but did not explore higher-dimensionality structures. To enrich the planned predictive validity comparisons with new lexicon-based inputs, our initial investigation went beyond this previous work to identify robust high-dimensionality structure in the 500-PDA. This novel structural investigation is detailed below in Preliminary Study 1.

The IPIP has been utilized to create public-domain counterparts of many originally proprietary personality inventories, but those counterparts are not examined here. Instead, we focus on the "real McCoys", the original inventories developed earlier in data outside the ESCS. The IPIP is used here, instead, for capturing a novel, relatively comprehensive assessment framework drawing on the cluster structure of many diverse non-IPIP scales administered to this sample, and for representing several derivatives of lexicon-based structures of personality. This novel framework in described below in Preliminary Study 2.

Before reporting the preliminary studies, it is pertinent to address why the present investigations focus on levels of personality variation that go beyond the Big Five and other few-factor models of personality, such as those associated with Eysenck, the Big Six or HEXACO model, and the Big Seven discussed by Saucier (1997). Previous studies on the predictive capacity of personality have compared the prediction afforded by a few broad domains of personality versus subcomponent or "facet" scales that are used as indicators of these broader aggregate or latent variables. Consistently, the facet scales are shown to add major increments to prediction (e.g., Ashton, 1998). This naturally leads to the suggestion that the broad domains be de-emphasized, or even ignored, in favor of the more numerous specific predictor variables. Thus, even if there were defensible consensus on one of these few-factor models of broad domains, studies of facet models would be a prime focus of research on personality and prediction.

There are reasons to be cautious about sole reliance on very broad domains of personality variation (such as the Big Five). First, such models are based on the lexicon but typically draw on only about the first 25% of variance in the data, as found in studies of the adjective lexicon. Aspects of the remaining 75% of the variance (much of which is not variation attributable to error) are drawn into the model only peripherally and inadvertently in measurement of subcomponents. Extracting a dozen or more additional factors appears to allow for an additional 10-15% of the overall variance in the adjective lexicon to be represented (Saucier & Iurino, 2019), which should widen predictive capacity.

Moreover, such broad domains have not been convincingly shown to represent universals, such that they would arise spontaneously in any cultural or linguistic context (without being constrained to appear by imposition of an imported measurement framework). Although many assume a structure like the Big Five is universal fact, this is a premature if not wholly unwarranted conclusion.

And finally, additional factors, more or less orthogonal to the basic five or six, have been identified. Saucier and Goldberg (1998); see also Paunonen & Jackson, 2000) detailed a number of specific sources of variation far enough beyond the Big Five that they empirically stand outside even very broadly measured versions of the Big Five. In other words, they cannot be facets or subcomponents of the Big Five. More recent work indicates that when one utilizes parallel analysis to guide answers to the "how many factors" question, and ceases to rely entirely on orthogonal varimax rotation methods, a relatively robust structure of some 20 factors can be recovered in multiple data-sets with North American participants (Saucier & Iurino, 2019); this structure draws in part on many of the peripheral sources of variation identified earlier by Saucier and Goldberg.

PRELIMINARY STUDY 1

As mentioned earlier, Saucier (1997) analyzed the structure of the 500-PDA (person-descriptive adjectives) at the few-factor level, with attention mainly to structures of one to seven factors. The focus was not only on the full set of 500 descriptors, which included many categories of descriptors (evaluative, social effect, temporary state, and physical and appearance descriptor terms) not ordinarily studied by personologists, but also on more conventional contents, including personality dispositions or on personality-plus-state dispositions. Saucier found that emergent structures depended to a considerable degree on variable selection. Although five-factor structures were relatively robust for narrower variable selections, a seven-factor structure was more appropriate for the full set of 500; these seven factors had partial resemblance to a Big Seven structure (e.g., a Negative Valence factor; Benet-Martinez & Waller, 2002) found in some other studies, but was in other respects different (e.g., included an Appeal/Attractiveness factor). The broad-factor-level structure of the 500-PDA, thoroughly investigated by Saucier (1997), is not further addressed here. Wood, Nye, and Saucier (2010) however drew on the full 500-PDA to identify a set of fine-grained individual-differences clusters with a wide range of contents, and the 61-cluster model they identified is further investigated in our main study.

Saucier and Iurino (2019) re-analyzed the structure of other, more conventionally narrow sets of English-language adjectives in English-speaking samples, relying on parallel analysis (and, with lesser effect on outcomes, the MAP method) to determine the number of factors. They found that reliance on varimax rotation tended to constrain the outcome to structures of roughly 12 factors or less, but that oblimin and equamax rotations identified some relatively robust structures in the range of 15 to 28 factors. The

DOI: 10.1002/per

convergences most strongly pointed toward 20 relatively independent factors, which were the most robust across variations in method (congruence between self- versus peer-report structures, the factor space identified by raw versus ipsatized data, factor-axis positioning found in orthogonal versus oblique rotations). Although robustness was less than what could be identified for few-factor structures (e.g., three to six factors), the difference was not great, and the demonstrated gain in predictive capacity and comprehensiveness appeared to outweigh the marginal loss in robustness.

The data-sets examined by Saucier and Iurino (2019) generally excluded many person-descriptor categories, these being exclusions similar to those in classifications by Allport and Odbert (1936) and Norman (1967), as well as Angleitner, Ostendorf, & John, 1990). According to these now fairly typical exclusions made in lexicon-based studies on personality attributes, evaluative terms (e.g., Likeable, Evil, Weird), temporary states (e.g., Joyful, Afraid, Tired), social roles and effects (e.g., Wealthy, Fascinating, Intimidating), and physical and appearance descriptors (e.g., Attractive, Slender, Short) were excluded. This begs the question of what high-dimensionality structure emerges in wider-inclusion variable selections like the 500-PDA. Filling that gap will enable an examination of the effect of wide versus narrow variable-selection strategies on predictive capacity, and filling that gap that is the focus of Preliminary Study 1.

Method

Participants. As described in detail by Saucier (1997), 700 participants in the Eugene-Springfield Community Sample (ESCS) provided self-reports on the 500-PDA Saucier (1997) also details a much smaller peer-report sample administered the 500-PDA, with 201 participants drawn from Western U.S. colleges or community colleges; this smaller set of data is utilized here exclusively for investigating how well the self-report structures generalize to a peer-report context. All the data-gathering research activities with the ESCS described here in later in this report were approved as exempt by the institutional review board of Oregon Research Institute.

Materials. The 500-PDA is detailed by Saucier (1997). It consists of those 500 adjectives that raters, both college students from California and a subsample of the ESCS, reliably identified as being the most frequently used. Materials included also 25 additional adjectives, mostly for the purpose of providing factor markers for the Big Five, but these additional terms are not studied here.

Analyses. A pre-registered analysis plan, based on the earlier applications of method by Saucier and Iurino (2019), was applied to 500-PDA data. Analyses entailed, in sequence: (a) parallel analysis (Horn, 1965) applied to both raw and ipsatized variants of the self-report 500-PDA data, to identify the largest number of factors to consider separately for each data-type (raw or ipsatized); (b) starting with that number of factors for each data-type, using principal components analysis as is standard for lexicon-based studies, by independent sequences using varimax, equamax, and oblimin (delta=0) rotations, working down successively to the largest number of factors that had all factors both sufficiently sized (at least three salient variables all with loadings over.30, one of those over.40, in absolute magnitude) and judged interpretable by at least one of the first two authors of this report; (c) setting that structure, specific to a data-type and rotation-method, as one of the candidate models, (d) comparing the six candidate models (e.g., raw-varimax, raw-oblimin, ipsatized-oblimin) on their robustness across three method variations (congruence between self- versus peer-report structures, the factor space identified by raw versus ipsatized data, factor-axis positioning found in orthogonal versus oblique rotations); and (e) selecting as best-supported whichever model, among the six, showed the most advantageous robustness across method-variations. Because robustness tends generally to decline as the number of factors increases, the ideal structure would have a minimal loss in robustness (compared to fewer-factor models) while allowing a gain in comprehensiveness with more factors, that is, the greatest gain in informativeness with the least loss in robustness.

One variation from the previous work of Saucier and Iurino pertains to step (a). Those authors employed the MAP (minimum average partial) procedure (Velicer, 1976) as a companion to parallel analysis to find a starting point for sequences of principal-components analyses. In three studies, the eventual outcome was found to be unaffected by the initial recommendations of MAP and thus attributable entirely to parallel analysis (except in the unusual case of structures based on clusters identified by Warren Norman, rather than the more usual analysis of single terms). They observed that in data with large number of single terms (rather than parcels or clusters), the MAP procedure often gave unrealistically high estimates of the number of factors, since no structures nearly that high met the criteria of having all factors sufficiently sized and interpretable. In other words, applications of MAP to data with large numbers of single terms led to solutions with numerous factors that were too small or uninterpretable for practical use. So, for this study, the application of MAP was dropped.

Results

In the main self-report data (N=700), parallel analysis indicated 30 factors for ipsatized data, and 23 factors for the raw (non-ipsatized) data. Sequences of principal-components analyses led to six candidate models. These were, for ipsatized data, structures of 13 varimax factors, 28 oblimin factors, and 30 equamax factors, along with, for raw data, structures of 13 varimax factors, 21 equamax factors, and 23 oblimin factors.

Table 1 provides the robustness-relevant coefficients comparing these six structural models. For reference, it also provides analogous coefficients for the seven-factor model based on ipsatized data, the model for the full 500-PDA best supported by Saucier (1997). Figure 1 graphically depicts the robustness coefficients and their mean, and Figure 2 depicts these on a scatterplot that better scales the increments of numbers of factors. In figures 1 and 2, the most desirably robust model would have the shortest distance to the upper

DOI: 10.1002/per

Table 1. Robustness Indices for the Seven-Factor and Six High-Dimensionality Models in 500-PDA

Candidate Model	Average Squared		
	Orthogonal-Oblique		Average Self-Peer Data
	Best-Match Correlation	Ipsatized vs. Original	Best-Match Congruence
7 factors (varimax)	.95	.78	.70
13 factors (varimax)	.75	.85	.61
13 factors (varimax)*	.78	.85	.58
21 factors (equamax)*	.83	.86	.71
23 factors (oblimin)*	.76	.85	.77
28 factors (oblimin)	.81	.87	.62
30 factors (equamax)	.64	.88	.55

Summed Proportion of Variance Explained

Note. N=861. Ipsatized data, including both self- and peer-ratings. ‘Summed Proportion of Variance Explained’ is the proportion of variance in one set accounted for by the other (derivable from the mean cross-loadings of individual factors in one set on the canonical variates derived from the other set). *based on raw rather than ipsatized data

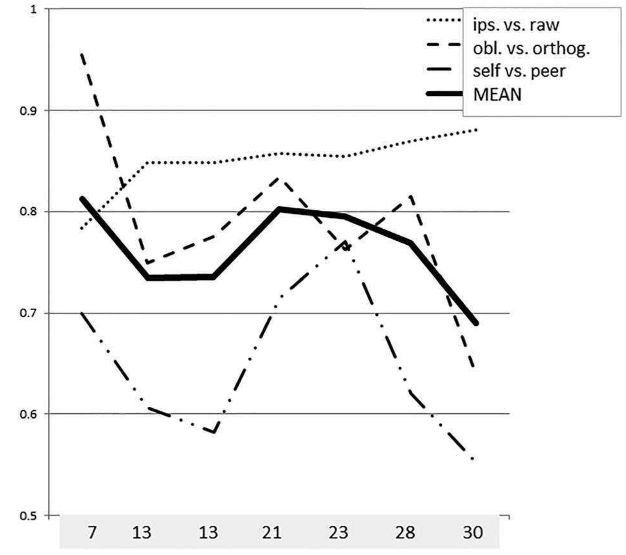

Figure 1. Robustness Indices for Six Candidate Models in 500-PDA

right (“northeast”) corner of the figure. Two models appear about equally desirable, having relatively high robustness for their level of comprehensiveness: 28 oblimin factors in ipsatized data, and 23 oblimin factors in raw data. The data to this point being ambiguous for selecting one of these models over the other, we retained both, and carried them over for further investigation in the main study.

Discussion

Table 2 provides a side-by-side comparison of essential core contents of the factors in the 23- and 28-factor models, with the convergent 20-factor model identified by Saucier and Iurino (2019). The 23- and 28-factor models have many similarities, but the 23-factor model has numerous factors – as is not unusual in factors from non-ipsatized data – that are not bipolar. The 28-factor model is more immediately comparable to the incoming 20-lexical-factor model (i.e., the Lexical-20 or even more abbreviated as the Lex-20, henceforth), which is sensible since both were originally identified in ipsatized data.

In the 500-PDA’s 28-factor model, 10 of the Lex-20 factors have readily identifiable counterparts, based on overlap in the defining terms. An 11th (Dominant/Demanding vs. Wishy-Washy) is a slightly more distant counterpart of a Lex-20 factor (Directness). The Truthfulness and Dependability factors from the Lex-20 are found combined into a single factor among the 28. The Courage factor from the Lex-20 divides into two factors among the 28 (Brave/Courageous and Daring/Adventurous), and the Lex-20 Affection/Emotionality factor divides also into two factors among the 28 (Romantic/Loving and Kind-hearted/Compassionate). The Sophistication factor from the Lex-20 has a marginal representation in the Prominent/Famous vs. Ordinary/Informal factor from the 28; undoubtedly the reason for the marginal representation is the poor representation of core Sophistication terms in the 500-PDA, terms such as Refined, Cultured, Dignified, and Polished being absent. As for the other four Lex-20 factors, they are entirely absent, for the same reason – a lack of very high-frequency adjective descriptors referencing content pertaining to Thrift, Guilelessness, Prejudice, and Reflectiveness.

What of the other 11 factors among the 28 factors, those which have little counterpart to the Lex-20? These mostly emphasize categories of content excluded by Allport and Odbert (1936) and Norman (1967) and most subsequent investigators. The Weird vs. Normal, Evil/Cruel, and Wonderful/Admirable factors emphasize evaluative terms. The Well vs. Disabled, Short vs. Tall, and Slender vs. Chubby factors emphasize physical characteristics, and the Pretty/Feminine factor has some emphasis on appearance characteristics. The Overworked vs. Unemployed factor emphasizes social-role characteristics, and the Rich/Secure vs. Lonely factor emphasizes social status. That leaves two factors as yet unaccounted for – Funny/Humorous and Attentive/Gracious/Clean. The first of these is found directly in some structures Saucier and Iurino (2019) identified, but not specifically in the Lex-20. The second is more unique to the present study, pertaining to some kind of propriety; it is noteworthy that two of the four most salient terms for the factor (Clean and Believable) are not even included in lexical variable selections analyzed by Saucier and Iurino (2019), and the other two (Attentive, Gracious) do not make consistent appearances there.

Based on these same data, Saucier (1997) argued that the structure of personal attributes is considerably impacted by variable selection, and these results reflect such impact. Thus, some nine or 10 of the 28 factors in the more informative model essentially could not have appeared in studies with more conventionally narrow selections of descriptors such as led to the Lex-20. And four or five of the Lex-20 factors fail to appear in the present data because their associated descriptors hardly appear at all in the variable selection. It would be difficult to argue that the present data refute the Lex-20 model, since to the extent variable selection allowed,

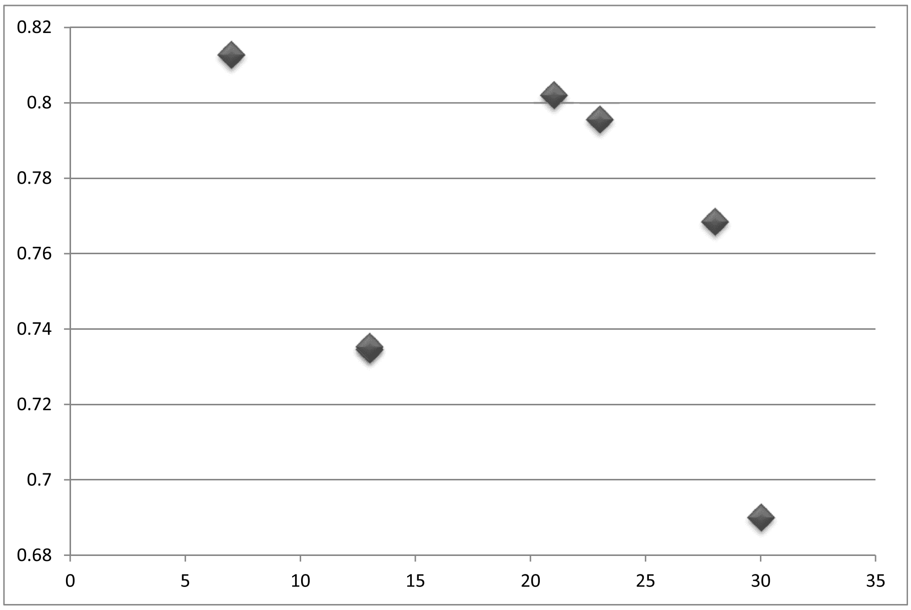

Figure 2. Robustness (y-axis) by Number of Factors (x-axis) [Colour figure can be viewed at wileyonlinelibrary.com]

Table 2. Comparison of Core Contents of 20, 23, and 28 Factor Models

23 factors, raw data	28 factors, ipsatized data	Lex-20 factors
Outgoing/Talkative vs. Shy	Outgoing vs. Shy/Quiet	1- Talkativeness (narrow Extraversion)
Happy/Glad/Joyful	Happy/Glad/Joyful	2- Enthusiasm/Positive Affect
Daring/Brave/Direct	Brave/Courageous Daring/Adventurous Demanding vs. Wishy-washy	3- (-)Fear-proneness 4- Directness/Firmness
Relaxed/Laid-back/Easygoing	Relaxed/Self-Assured vs. Anxious Prominent/Famous vs. Informal	5- (-)Anxiety-proneness 6- Sophistication
Intelligent/Smart vs. Uneducated	Intelligent/Smart vs. Ignorant	7- Knowledge/Intellect 8- Reflectiveness
Creative/Talented/Imaginative	Creative/Talented/Artistic	9- Originality
Loving/Warm-hearted vs. Cold	Romantic/Passionate/Loving	10- Emotionality 11- Prejudice
Crabby/Irritable/Short-tempered	Crabby/Irritable vs. Patient	12- Stubborn/Temperamental
Cruel/Terrible/Bad	Radical vs. Proper/Polite Cruel/Evil/Terrible	13- Rudeness/Cruelty
Selfish/Greedy/Self-centered	Snobbish/Self-centered	14- Egotism (vs. Humility) 15- Guile/Cunning
Responsible/Honest/Trustworthy	Honest/Trustworthy/Reliable	16- Dishonesty 17-Dependability
Organized/Neat vs. Sloppy	Organized vs. Messy/Disorganized	18- Order/Organization 19- Practicality/Thrift
Conservative/Traditional	Conservative vs. Open-minded Clean/Attentive/Gracious	20- Conventionality
Funny/Humorous	Funny/Humorous Kind-hearted/Compassionate	
Pretty/Feminine vs. Masculine	Pretty/Feminine vs. Masculine	
Secure vs. Lonely/Lonesome	Secure/Rich vs. Lonely	
Weird/Strange vs. Normal	Weird/Strange vs. Normal	
Terrific/Wonderful/Great	Wonderful/Admirable vs. Awful	
Young/Employed vs. Elderly Busy/Influential/Overworked	Employed/Overworked vs. Unemployed	
Well/Healthy vs. Tired/Disabled	Well vs. Disabled/Handicapped	
Slim/Slender vs. Chubby	Slim/Slender vs. Chubby	
Short/Little vs. Tall	Short/Little vs. Tall	

the various Lex-20 factors did tend to appear though occasionally in combined/paired or bifurcated form. About half of the Lex-20 factors – counting here Talkativeness, Enthusiasm/Positivity, Politeness, Patience vs. Bad Temper, Orderliness/Meticulousness, Relaxation vs. Anxiety, Affection/Emotionality, Knowledge/Intellect, Creativity, and Conventionality – appear in highly similar form despite the radical difference in the nature of the overall variable selection.

Accordingly, both the high-dimensionality Lex-20 model from narrower variable-selection studies, and the two most promising high-dimensionality lexicon-based models from the present ESCS data, are carried forward to the main study. The main study might say something about the comparative merits of these three candidate models, from the standpoint of comparative prediction.

PRELIMINARY STUDY 2

As described earlier, the co-occurrence of the IPIP and a wide variety of personality inventories in the context of the Eugene-Springfield Community Sample (ESCS) enables one directly to develop and faithfully represent a new kind of framework -- a novel, relatively comprehensive assessment framework drawing on the cluster structure of many diverse non-IPIP scales administered to this sample. Our intent was to develop a fine-grained structure of many relatively small components of variation present in instruments administered to the ESCS, without any necessary reference to a higher-order structure, and we judged hierarchical cluster analysis to be the method of choice for this desideratum. The goal was to fit as many non-IPIP personality-relevant variables as possible into meaningful clusters, and then develop a small set of IPIP items (ideally, four items for each) as a standard assessment for these clusters. The four-item standard was influenced by the similar-length homogenous item-composites found in the Hogan Personality Inventory (Hogan & Hogan, 1995). With a general aim of having questionnaire similar in length to a widely referenced personality inventory with 240 items (the NEO-PI-R; Costa & McCrae, 1992), the desire was to identify roughly 60 clusters.

Method

Participants. Members of the ESCS provided responses to a wide variety of non-IPIP instruments in the period from 1993 to 2006. The sample sizes for the various questionnaire-packages in which these were included vary, but are referenced in the ESCS technical report on the ESCS (Goldberg & Saucier, 2016). The sample sizes range from a high of 856 for the NEO-PI-R in 1994 down to a low of 663 for several scales (Narcissism, Sensitivity to Reward and Punishment) administered in 2006. The common sample that completed all the measures referenced consisted of 352 participants, and that sample was used for the key derivation cluster analysis.

Materials. A wide variety of variables administered to the ESCS that were included in the analysis, and we included all that could be judged a personality variable. These included all non-IPIP multiscale inventories used later in the main study, plus non-IPIP scales from 14 other sources (e.g., self-esteem, self-monitoring, adult attachment). The VIA-IS scales were included; it is worth noting that the VIA-IS items all later became part of a continually expanding IPIP (online the larger set of virtues-relevant items that includes the VIA-IS are coded as "V" variables in the IPIP). In all, 257 variables were analyzed.

Analyses. The hierarchical cluster analysis used Pearson correlations as the distance measure. Our reading of the literature suggested that between-groups linkage is the most commonly used clustering method. But in our experience another method, complete linkage (i.e., furthest-neighbor analysis) is a strong competitor to between-groups linkage, as it tends to yield relatively equally sized clusters to at least as great an extent as between-groups linkage does. Our initial analyses employed both methods. But we found the complete-linkage results tended to yield fewer singletons (unclustered variables), which was a desirable outcome given our goals, and therefore we relied on this method by preference. Variables were grouped into clusters when they appeared adjacent to one another on the dendrogram and linked hierarchically (thus, adjacent variables not linked hierarchically were placed on separate clusters). Clusters were separated from one another within the dendrogram if they diverged above a certain fixed level (inside the 10th level of agglomeration in the dendrogram), which tended to keep separate scales whose correlations fell well under.20 in magnitude. This level was chosen because it appeared to allow for a manageable total number of clusters (i.e., in the 50-75 cluster range), with overlarge clusters (more than six constituent scales) divided into subsidiary clusters so that no cluster drew on more than six scales.

When cluster-analyzing a set of variables with many negative intercorrelations it is often advisable to make an adaptation allowing negatively correlating variables to form a cluster. The obvious rational method for allowing this to occur is to include in the analysis, for each variable, both the original score and a reflected score that correlates -1.00 with the original score. With this adaptation, the cluster tree produces two mirrored halves, and one need only examine one of the two identical halves. The adaptation just described was applied in this analysis.

Results

The complete-linkage dendrogram, based on 514 variables (257 scales plus the reflected version of each), is very long and unsuitable to be a table or figure in a scientific report, but we present the eventually-retained clusters, in roughly the order in which they initially appeared in the dendrogram, in Table 3. There were initially 68 clusters consisting of two or more scales, and in addition seven singleton clusters were provisionally retained, in case they could be found to capture useful additional variance beyond the multiple-scale clusters. An initial benchmark for each cluster was derived by producing component scores

DOI: 10.1002/per

Table 3. Clusters in the Integrative Personality Questionnaire

Name	L-20	IPQ	items	alpha	Viir	Name	L-20	IPQ	items	alpha	Viir
Social sensation seeker	1	2	4	.61	.005	Spiritual beliefs	10	39	4	.85	007
Humor	1	1	4	.53	.067	Excitable	10	47	2	.52	--
Sociality	1	33	4	.50	.003	Love of beauty	10	14	4	.45	.006
Open and expressive	1	16	4	.72	.009	Empathetic/sensitive	10	17	4	.67	.020
Social confidence	1	53	4	.78	.002	Good listener	11	18	2	.47	--
Social grace	1	15	4	.60	.006	Equality/compassion	11	20	4	.51	.001
Gregarious	1	52	4	.75	.004	(-) Violently avenging	12	70	2	.13	--
Friendliness	1	34	4	.73	.007	Tolerant	12	22	4	.49	.005
Trusting	2	66	4	.67	.003	Cool temper	12	26	4	.73	.004
Happy	2	56	4	.61	.004	Patient	12	25	2	.66	--
Well-being	2	57	4	.73	.004	(-) Vengeful	12	30	4	.57	.001
Future oriented	2	58	4	.72	.030	Caring/soft-hearted	13	19	4	.48	.002
Robust	2	54	4	.78	.002	(-) Boorish	14	4	4	.65	.004
(-) Thin skinned	2	23	3	.40	.003	Humble	14	28	4	.62	.006
Physically robust/strong	2	61	4	.69	.001	(-) Social striver	14	5	4	.61	.001
(-) Belief in luck/fate	2	49	4	.58	.011	(-) Lover of luxury	15	29	4	.46	.002
Self-belief	3	60	4	.70	.002	(-) Acts to impress	15	67	4	.62	.005
Fearless	3	63	4	.69	.006	(-) Deceptive manipulator	15	6	4	.56	.001
Thrill seeker (extreme sports)	3	8	4	.72	.013	(-) Cheater	16	27	4	.62	.003
Mechanically inclined	3	7	4	.64	.019	Unlikely moral virtues	16	43	4	.52	.004
Numerically inclined	3	32	3	.60	.009	Grateful/respectful	17	21	4	.52	.003
Indifferent to opinion	3	36	4	.60	.012	Belief in effort	17	74	3	.49	0
Self-confident vs. indecisive	4	59	4	.62	.001	Hard-working	18	44	4	.54	.006
Unworried	5	62	4	.76	.006	Tidy	18	41	4	.66	.013
(-) Sensory delusions	5	64	4	.69	.004	Goal-motivated	18	55	4	.67	.004
Concerned with appearance	6	37	4	.52	0	Detail-oriented	18	42	4	.71	.006
Wise, good judgement	7	51	4	.67	.010	Self-controlled	18	45	4	.63	.007
Serious reader	7	13	4	.65	.025	Deliberate	18	40	4	.61	.012
Reader	7	48	4	.67	.022	Financially prudent	19	72	4	.67	.004
Academically inclined	7	31	4	.70	.028	Respect for authority	20	38	4	.53	.006
Self-analytical	8	46	4	.69	.006	(-) Spiritual experiences	20	12	4	.69	.019
Unconventional thinker	8	11	4	.66	.006	(-) Thrill seeker (reckless)	20	3	4	.62	.001
Imaginative	9	10	4	.73	.005	Creature of habit	20	9	4	.70	.004

Note. L-20 = factor from Lex-20 structure (see Table 2) with which scale has strongest association. IPQ = original scale number from cluster analysis. Viir = Variance interitem r. Scales are ordered based on the Lex-20 scale with which each is most associated.

from the first unrotated principal component for the combined constituent scales in the cluster.

In order to be retained ultimately and carried forward into the main study as what we called an Integrative Personality Questionnaire (IPQ), the clusters had to be measurable using the item-resources of IPIP. To qualify as a potential cluster-marker item, an IPIP item had to have its highest correlation with that cluster, and no other, and that correlation had to exceed.25 in absolute magnitude. We employed a mechanistically rule-based, and pre-registered, procedure for selecting items. First, a set of eight initial IPIP items was selected – preferably four having positive and four having negative correlations with the cluster – with preference for those having a correlation with the cluster that doubled in value the correlation with any other cluster (in other words, univocal on the cluster). Items with that ratio being at least 1.5:1 (rather than 2:1) got secondary preference, but if insufficient numbers of items were obtained without meeting even this criteria items were selected if the ratio simply exceeded 1:1. When more than four items qualified at any of these "univocality-ratio" levels, preference was given to those having higher correlations with the cluster.

Once the initial pool of eight items was selected, the correlation matrix among these items was scrutinized with the aim of removing outlier (overly high and especially overly low) correlations from the matrix, resulting in as much as possible a set of four items with relatively homogenous intercorrelations (and thus, relatively unidimensional). Specifically, from a set of 28 intercorrelations in the matrix formed by eight items, the four highest and eight lowest correlations were identified, and those items which participated in the greatest number of these outlier correlations were targeted for exclusion. The one exception to this rule: If an item was in a class with a higher univocality ratio (see previous paragraph) than other items, it was preferenced for retention regardless of the outlier-correlation count. A further rule regarded ties: If multiple items qualified for the last available spot on the scale, that item was preferred which would tend to increase the variance of the item means (a criterion that tends to enhance equidiscrimination at all levels of the trait from low to high). It is worth noting this item selection approach does not maximize internal-consistency reliability, while it does tend to maximize unidimensionality. The low variance in the inter-item correlations for these cluster scales, evident in Table 3 – most usually under .01 meaning the standard deviation of the correlations was under.10 – gives some indication of this tendency toward unidimensionality. More

DOI: 10.1002/per

detailed information, including the IPQ items, is found in a supplementary table.

Although reliability was not the most important goal, in fact all but one cluster-scale has alpha over.45, and most are quite a bit higher. The typical number of items per cluster is four, but some clusters had fewer than four qualifying items. The total number of scales was 66, measured by way of 262 IPIP items. Only four of the seven singleton scales survived the thresholds of measurability using IPIP described above; the other 62 cluster scales derived ultimately from combinations of non-IPIP scales.

In Table 3, the IPQ's cluster-based scales are listed in a different order than cluster-analysis dendrogram listed them. We placed them in a content-order corresponding to the order of the Lex-20 constructs in Table 2. The order was created by correlating IPQ scales with Lex-20 adjective scales to form 20 subgroups, then ordering the subgroups based on their relative level of correlation with the next highest or next lowest Lex-20 scale on the Table-2 list. The ordering of the Lex-20 scales follows that in the original presentation by Saucier and Iurino (2019), which was derived based mainly on how the 20 scales related to an underlying Big Two (bivariate) model. The arc of personality descriptions in either case (Table 3, or Table 2's right column) involves a spectrum from impulse-expression to aspects of intellect and emotional and moral aspects of personality, culminating with attributes emphasizing self-control.

Discussion

Preliminary Study 2 was essentially a scale-construction exercise. The derivation of an integrative personality questionnaire based on cluster-analysis of personality-relevant scales administered to the ESCS was highly relevant to the goals of our overall investigation, as it would provide a useful complement and comparison to other assessment platforms. Such a cluster-analysis approach affords a structural framework likely to go, at least in part, beyond the Big Five, and indeed has no reliance on factor analysis, which may tend to produce mainly large, coarse-grained aggregations of variables. Preliminary Study 1 yielded no structure with more than 28 factors, and thus 28 variables available for analysis, and this study yielded a framework with over twice that number. Similarly, Wood, Nye, and Saucier (2010) applied hierarchical cluster analysis to the 500-PDA (plus the 501st through 504th most frequent adjectives, which were also included in the 1995 data) and in a somewhat similar way arrived at 61 clusters constituting the Inventory of Individual Differences in the Lexicon (IIDL), although those clusters were based on person-descriptive adjectives and not questionnaire scales.

The IIDL is based on pairs of adjective items, whereas the present Integrative Personality Questionnaire (IPQ) is based on (mostly) quartets of IPIP items. But both platforms enable a viewing angle on what personality structure might look like with not five or six, and not 20 or 23 or 28, but instead over 60 constituent variables. In other words, either one (with its own distinct basis) provides a more fine-grained version of high-dimensionality personality structure. In our main study we compare such fine-grained versions with factor-based high-dimensionality representations, and compare both these with facet representations from conventional facet-and-domain organizations in prominent personality inventories.

MAIN STUDY

The main study compares 11 structures ultimately derived from outside the ESCS as well as IPIP, with respect to predictive validity, but also draws into the comparison those instruments referenced or derived in the two preliminary studies. In the case of lexicon-derived structures, each structure is represented by two alternative methods (adjective-scale vs. IPIP-scale, or adjective-scale versus factor scores). A total of 21 sets of scales (including a couple of factor-score indices) were included.

The principal questions for the main study are four. Which set of scales will predict the best, overall? How is predictive capacity related to the derivation-source for the model (whether the lexicon or some expert-selection basis)? Does an increase in the number of distinct predictor variables produce a dependable increase in predictive capacity? Does it help to have more items instead of fewer items informing the predictor-variable scores? And how does another related indicator of length, the number of words in the questionnaire, bear on predictive capacity? Overall, results might indicate not only which inventory-platforms are more or less strong predictively, but also the general outlines of a personality-assessment instrument that is most predictively efficient.

Predictive efficiency would be, ideally, the maximum amount of variance accounted for, in the criterion/outcome variables, with the minimum stimuli, the minimum number of items and words – the minimum "item cost" (Yarkoni, 2009). Apart from item cost – indicated by the length of items and words the assessment requires – it could potentially be desirable to have the minimum number of predictor variables, all else equal, because the smaller number of variables in the model the more parsimonious, and the easier the model will be to explain (theoretically or otherwise). It should be easier to make a theory about, and arrive at a precise interpretation of, one variable than about a set of 60 variables. However, the emphasis in this study is on maximizing prediction and not on maximizing parsimony or theory-relevance.

Nonetheless, we adopt some minimum considerations with respect to construct validity. Cronbach (1990) defined validation as "inquiry into the soundness of interpretations proposed for the scores from a test" (p. 145), implying that the end-point of validity is a sound interpretation of what is being measured. Similarly, Messick (1988, p. 39) defined "strong construct validation" in terms of "evidence discounting plausible rival hypotheses" for how to interpret scores. There are difficulties in arriving at sound interpretations of the meaning of variables measured by a single item, for two reasons. First, one cannot introduce balanced-keying which would allow for control over acquiescence bias; therefore, a single-item variable will be inevitably haunted by the rival hypothesis that scores reflect (to some noise-producing degree) differences in indiscriminate yea-saying versus nay-saying. Second, the aggregation of multiple items into a score enables one to gauge the degree to which various contents (including distracting elements like accidentally similar item framing) are contributing to the score and its internally consistent core, and one cannot estimate internal consistency from a single item; especially with complexly worded single items, it can be difficult to determine which aspects of the wording or framing are contributing most to response variance. Based on these considerations, we examined here only inventories where variables are measured with multiple items each, and thus have a higher ceiling with respect to the interpretability and thus construct validity of the constituent variables. Interpretable scores are useful not only for purposes of validity, but also as a necessary scaffold for building theory.

METHOD

Participants. As in the preliminary studies, the participants were members of the Eugene-Springfield Community Sample. The analyses detailed below entailed pairing 12 criterion variables in turn with each of 21 “assessment platforms.” Had we relied on only that subset of participants who had values for all 12 criterion variables and all 21 assessment platforms, the effective (listwise) sample size would have been overly small. To maximize precision and statistical power, we adopted a pairwise approach, using for each analysis the maximum sample size available for that combination of variables. These sample sizes ranged from 368 to 701, except for the analyses involving one of our outcome variables (BMI, see below), for which systematically fewer data points were available, so that sample sizes ranged from 228 to 396. Generally, variables that relied on later-administered questionnaires (e.g., 2006 rather than 1994), and on IPIP items had lower sample sizes. IPIP-item scales have systematically lower sample sizes because IPIP items were administered at various diverse time-points and only those participants who overlapped in having data from those data points would be represented. There was no missing-value imputation carried on for this study, beyond what has already been incorporated in standard publicly-available data for the ESCS.

Materials. There were 12 criterion variables, each of which was measured on a continuous (not categorical, binary, or strictly ordinal) scale. The predictor variables were naturally grouped into 21 sets. These involve 11 sets representing personality assessment structural models developed originally outside the ESCS, and 10 sets for which the model or measure was derived within the ESCS. The 11 models imported from outside are described first.

1. Multidimensional Personality Questionnaire (MPQ: Tellegen & Waller, 2008). We analyzed 12 variables based on 276 administered items including 3374 words.
2. Jackson Personality Inventory – Revised (JPI-R; Jackson, 1994). We analyzed 15 variables based on 300 items including 3572 words.
3. 16PF (Conn & Riek, 1994). We analyzed 16 variables based on 185 items including approximately 2979 words.
4. Six-Factor Personality Questionnaire (6FPQ; Jackson, Paunonen, & Tremblay, 2000). We analyzed 18 variables based on 108 items including 1200 words.
5. HEXACO-PI (Lee & Ashton, 2004). We analyzed 24 variables based on 192 items including 2292 words.
6. Values in Action – Inventory of Strengths (VIA-IS; Peterson & Seligman, 2004). We analyzed 24 variables based on 214 items including 1614 words. One caution with this instrument is that the reduction from the more standard set of 240 items drew on item-performance in the current data (based on archival records). The fact that some of the instruments below using IPIP items may draw on some of these same 214 items (which were later incorporated in IPIP) has no effect on these predictive-validity analyses; it would only serve to increase the correlation between the VIA-IS scores and these other predictor sets.
7. NEO Personality Inventory – Revised (NEO-PI-R; Costa & McCrae, 1992). We analyzed 30 variables based on 240 items including 2262 words.
8. Temperament and Character Inventory (TCI; Cloninger, Przybeck, Svrakic, & Wetzel, 1994). We analyzed 31 variables based on 295 items including 4147 words.
9. California Psychological Inventory (CPI; Gough, 1996). We analyzed 36 variables based on the 462 items in the CPI, which include 5547 words.
10. Hogan Personality Inventory (HPI; Hogan & Hogan, 1992). We analyzed 44 variables based on 206 items (on an HPI version with a 1990 copyright date) including 1580 words.
11. Adjective Scales for Lex-20. This set of 95 adjectives was derived based on North American college-student-sample data as an exportable marker-set by Saucier and Iurino (2019). All were administered at some point to the ESCS, with a 1-to-7 response scale, or a 9-point scale rescaled to that. There were 20 variables, 95 single-word items, thus 95 words.
12. Big Five Aspect Scales (BFAS; DeYoung, Quilty, & Peterson, 2007). We analyzed 10 variables based on 98 items including 451 words. The BFAS in total has 100 items, but two of those items were not administered to the ESCS. Like all the variable-sets described subsequently, the BFAS was developed using ESCS data.
13. Yarkoni’s 30 NEO facet scales (here, NEO-Yarkoni). Yarkoni (2009) derived a set of 181 items to use for scoring a variety of inventories that had been administered to the ESCS. Here we use only the 108-item set employed for capturing NEO-PI-R facets. Notably, numerous items are included in more than one facet scale, so the multiple-regression analysis here will be partialing out much collinearity among predictor variables. We analyzed 30 variables based on 108 items including 590 words.
14. AB5C-IPIP. The AB5C (abridged Big-Five circumplex) model was originally developed using adjective variables in college-student samples (Hofstee, De Raad, & Goldberg, 1992). These adjectives were not administered at

DOI: 10.1002/per

one uniform time to the ESCS. IPIP versions of the AB5C developed by Goldberg are now the more standard assessment format. We analyzed 45 variables based on 486 items including 2215 words.

15 Inventory of Individual Differences in the Lexicon (IIDL; Wood, Nye, & Saucier, 2010). We analyzed 61 variables based on 122 items, 122 words.

16 An Integrative Personality Questionnaire (IPQ). This model, described above in Preliminary Study 2, involves 66 variables, but one of these variables (cluster 64) was held out because it was based in part on the somatoform dissociation scale (criterion 11 below). The remaining 65 variables draw on 258 IPIP items using 1571 words.

17 IPIP Scales for Lex-20. The 95 adjective markers were correlated with IPIP items (using the same item-selection method as Preliminary-Study 2) to arrive at four-item scales for each factor (coefficient Alpha from.52 to.77), thus 20 variables, 80 items, 409 words.

18 Adjective Scales for 23 lexical factors. Table 2 provides a guide to predominant content, all from the 500-PDA, from Preliminary Study 1; 23 variables, 115 items, 115 words.

19 Adjective Scales for 28 lexical factors, also in Table 2; 28 variables, 112 items, 112 words.

20 Factor scores for 23 lexical factors. Here the 23 variables are factor-scores derived from the 500-PDA; 23 variables based on 500 items, 500 words.

21 Factor scores for 28 lexical factors. Here the 28 variables are factor-scores derived from the 500-PDA; 28 variables, 500 items, 500 words.

The numbers of variables and items in the models, as described above, are exact. The number of words involved in the items measuring each model are, we believe, essentially accurate. For most models, we had digital lists and could machine-count. For sets 1-4, 7, and 9-10 above, which are all commercially published propriety inventories, we had available only hard-copy materials marked with copyright notice (and no electronic files listing item wordings). Our word-counts for these proprietary instruments are based on hand-counting. Out of respect for U.S. copyright law, we refrained from any photocopying or digital scanning of these materials. If incorrect, our hand-count estimates are likely off by only a few digits, and such a difference (i.e., about 1/10 of a percent) would not meaningfully influence results.

The 12 criterion variables were as follows.

1 Risk-avoidant Health Practices. A 10-item subset of a Health Practices Questionnaire (HPQ) developed by Goldberg based on items from earlier health-related inventories, as described in Goldberg and Saucier (2016), administered in 1995. An example item: “I cross busy streets in the middle of the block.”

2 Good Health Practices. A different 12-item subset of the aforementioned HPQ. Example item: “I eat a balanced diet.”

3 Health Concerns. A third, 15-item subset of the HPQ. Example item: “I gather information on things that affect my health.”

4 Peak Tobacco Use-Level, in Lifetime (to 2006). Question elicited how many (packs of) cigarettes respondent was smoking per day, at the point of peak tobacco use during their life (non-smokers had a zero value on the variable). Selected as a potential indicator of tobacco-addiction tendencies.

5 Peak Alcohol Use-Level, in Lifetime (to 2006). Question elicited how many alcoholic drinks respondent was consuming per week, at the point of peak alcohol consumption during their life (non-drinkers had a zero value on the variable). Selected as a potential indicator of binge- or heavy-drinking tendencies.

6 Body-Mass Index (BMI). The BMI index used here, based on self-reported height and weight, was already adjusted for sex of participants, so step 1 coefficients are minimal.

7 Satisfaction with Life Scale (SWLS; Pavot & Diener, 1993). Five-item happiness measure administered in 2001.

8 Depression symptoms as captured by Center for Epidemiological Studies Depression Scale (CES-D; Radloff, 1977). 24 items, expansion on CES-D, administered in 1997.

9 Mental-Health Conditions. An item assessing how many of four different conditions one had ever been diagnosed with (by self or a professional) or treated for: anxiety/panic, depression, bipolar, and schizophrenic disorders.

10 Fears Questionnaire (Marks & Mathews, 1979). Respondents asked to indicate how many of 25 types of fears (phobia symptoms) they experience. Administered in 2006.

11 Somatoform Dissociation Questionnaire (Nijenhuis, Spinhoven, van Dyck, van der Hart, & Vanderlinden, 1997). Administered in 2006.

12 Aggregate of Self-Report Psychopathy Items (SRP; Paulhus, Neumann, & Hare, in press). These were 12 SRP items that were not presented in IPIP format, and thus cannot be confounded with measures based on IPIP items.

Analyses. Each of the 21 sets of variables was employed as one set of predictors (independent variables) in a linear multiple regression with each of the 12 criterion variables (as dependent variable). Because there is some tendency toward inflation in multiple correlations (and the related R-squared values) as the number of predictors increase, we relied on adjusted (shrunken) R-squared values as one prediction coefficient. The regression was hierarchical, in that age and gender (male, female) was entered at a first step, and one of 21 sets of variables at a second step. Accordingly, this prediction coefficient was not the adjusted R-squared after step 2, but rather the change in adjusted R-squared from adding the step-2 predictors. Due to formulaic shrinkage in R-squared, occasionally the adjusted R-squared after step 1 was a slightly negative coefficient (e.g., -.001), but these are reported as exactly zero since in reality R-squared cannot be less than zero.

As an alternative way of inflation-correcting R-squared values, we used *k*-fold cross-validation. There is debate

DOI: 10.1002/per

about which number of folds is optimal in general, or for various data-analysis situations (as noted by Afendras & Markatou, 2019), but folds in the three to 10 range are common. On an *a priori* basis we selected the 5-fold cross-validation as our primary method because, had we been creating new measurement models and testing them with sample sizes as small as 360, any division into five would allow cases always to outnumber predictors (which here reached as high as 67). This selection was somewhat arbitrary, since here we were not generating novel measurement models, but simply taking regression formulae from a training set and examining their outcome in test sets. So we examined also other variants (3-fold, 7-fold, and 10-fold) and found that they generated results that converged very highly with the 5-fold results. For archival purposes (at least) we retained descriptive statistics, correlation matrices, and slope coefficients for each variable in each analysis, but for sake of reasonable economy we do not report these here.

To observe the ways in which differing numbers of variables, items, and words in the predictor models might affect results, we report how the (21x12=) 252 adjusted-R-squared values, or alternatively, 5-fold-cross-validated values (based on 100 repetitions), mapped onto how many variables, items, and words are involved in producing the scores in each of the 21 sets of variables.

Results

Because the adjusted-R-squared values tended overall to be higher than the 5-fold-cross-validated values, in the interests of caution we gave priority to the latter. In tables 4 through 7, the 5-fold-cross-validation values (change in R-squared after prediction of age and gender) are listed first, with the corresponding change in adjusted-R-squared in parentheses, and the sample size below that. We used the maximum sample size available for each combination of variables.

Table 4 presents prediction coefficients for the first 11 (non-ESCS-derived) models, in relation to the six health-related criterion variables. The highest cross-validated prediction coefficients tended to come with the VIA, TCI, and JPI. With conventional adjusted R-squared, the NEO, TCI, and VIA had the strongest results.

Table 5 presents coefficients for the same 11 models, and the six psychopathology-related criteria. Here, the MPQ dominated on most of the criteria, using cross-validation. With conventional adjusted R-squared, the CPI and MPQ had the strongest prediction returns.

Table 6 provides coefficients for the remaining 10 models, all derived in ESCS data and thus with some particular hazard of inflation in prediction coefficients based on capitalization on chance elements in ESCS data. Factor-score

Table 4. Cross-Validated (and Adjusted) R-squared Values for Health-Related Criteria: Models Developed Previously Outside This Data

Inventory	Risk-avoidant HP	Good HP	Health Concerns	Peak Tobacco Frequency	Peak Alcohol Frequency	Body-Mass Index (BMI)
MPQ	.17 (.19) 583	.08 (.11) 583	.12 (.14) 583	.02 (.04) 596	.08 (.10) 597	.00 (.01) 396
JPI	.21 (.23) 575	.10 (.13) 575	.09 (.11) 575	.04 (.06) 590	.10 (.12) 591	.00 (.02) 385
16PF	.20 (.22) 567	.06 (.09) 567	.06 (.09) 567	.02 (.04) 535	.06 (.09) 536	.00 (.02) 349
6FPQ	.09 (.12) 560	.06 (.09) 560	.09 (.12) 560	.02 (.04) 570	.04 (.06) 570	.00 (.01) 374
HEXACO-PI	.17 (.21) 577	.15 (.19) 577	.06 (.09) 577	.02 (.04) 638	.03 (.07) 639	.04 (.08) 377
VIA-IS	.21 (.25) 539	.17 (.21) 539	.09 (.14) 539	.02 (.04) 617	.06 (.09) 618	.16 (.22) 347
NEO-PI-R	.20 (.24) 635	.08 (.13) 635	.09 (.14) 635	.04 (.08) 585	.06 (.11) 587	.01 (.04) 381
TCI	.24 (.28) 593	.08 (.11) 593	.12 (.14) 593	.02 (.04) 568	.08 (.10) 568	.00 (.01) 371
CPI	.18 (.22) 620	.07 (.12) 620	.06 (.11) 620	.02 (.05) 568	.05 (.10) 570	.00 (.05) 369
HPI	.15 (.20) 597	.05 (.11) 597	.03 (.08) 597	.02 (.07) 569	.00 (.05) 571	.04 (.10) 373
Adjectives, 20 Lex. Factors	.16 (.20) 488	.03 (.07) 488	.05 (.10) 488	.01 (.03) 467	.05 (.09) 467	.03 (.06) 286

Note. 5-fold cross-validated change in R^2 (after accounting for age and gender) with conventional adjusted R^2 change in parentheses, N for analyses below. HP – Health Practices.

Table 5. Cross-Validated (and Adjusted) R-squared for Psychopathology-Related Criteria: Models Developed Previously Outside This Data

Inventory	Life-Satisfaction		CES Depression		Mental Health		Fear Questionnaire		Somatoform Dissociation		Self-Report Psychopathy	
MPQ	.30	(.32) 645	.28	(.30) 691	.12	(.14) 662	.14	(.16) 597	.23	(.26) 674	.07	(.09) 588
JPI	.15	(.17) 406	.18	(.21) 405	.09	(.11) 407	.15	(.17) 408	.13	(.16) 404	.09	(.11) 400
16PF	.19	(.22) 575	.22	(.25) 639	.10	(.12) 586	.13	(.16) 537	.13	(.16) 593	.02	(.05) 528
6FPQ	.07	(.10) 557	.08	(.11) 617	.03	(.06) 568	.07	(.10) 516	.08	(.10) 582	.04	(.07) 508
HEXACO-PI	.22	(.26) 663	.17	(.21) 678	.09	(.12) 701	.16	(.19) 639	.14	(.18) 675	.08	(.11) 631
VIA-IS	.30	(.34) 625	.21	(.24) 630	.10	(.14) 665	.11	(.14) 618	.17	(.20) 626	.06	(.10) 609
NEO-PI-R	.19	(.24) 626	.19	(.23) 696	.09	(.13) 649	.10	(.14) 586	.12	(.16) 654	.06	(.10) 578
TCI	.18	(.23) 619	.22	(.27) 683	.08	(.12) 628	.11	(.16) 569	.22	(.26) 637	.03	(.08) 560
CPI	.21	(.27) 603	.23	(.28) 669	.09	(.14) 616	.19	(.24) 569	.19	(.25) 624	.12	(.17) 561
HPI	.25	(.31) 621	.27	(.33) 686	.07	(.13) 630	.11	(.18) 570	.15	(.21) 641	.06	(.12) 561
Adjectives, 20 Lex. Factors	.12	(.17) 516	.12	(.16) 516	.06	(.09) 509	.09	(.13) 467	.09	(.13) 507	.06	(.09) 459

Note. 5-fold cross-validated change in R^2 (after accounting for age and gender) with conventional adjusted R^2 change in parentheses, N for analyses below

indices from the 500-PDA (preliminary study 1), particularly the 23 factor structure, dominated cross-validated prediction in this table. But with conventional adjusted R-squared, the IPQ showed the highest prediction value for most of these health-related criteria.

Table 7 has coefficients for these latter 10 models for psychopathology criteria. With cross-validated prediction, the 23-factor-score and IPQ models performed strongest. But with conventional adjusted R-squared, the IPQ again had the highest prediction value for these criteria.

Figure 3 depicts the prediction coefficients for each inventory in boxplot format, showing the mean (and standard deviation units of the dispersion) of these coefficients, and ordered by the size of the unadjusted, non-cross-validated change in R-square values. Also included for a benchmark comparison are the coefficients for the five NEO-PI-R domains (based on 240 items, designated as NEOAC), which show up on the low end of the inventories' distribution. It can be seen that method -- use of *k*-fold cross-validation versus adjusting for the number of predictor variables -- substantially affects the relative size of the coefficients.

One problem with coefficients in Tables 4 through 7 is that the highest values in the tables reflect virtual tautologies. They depict dramatically higher prediction of BMI from certain models, which happen to be the 500-PDA-based models that include a variable assessing how Slender versus Chubby a person is – an obvious way of informally assessing BMI. Such tautologies would distort the between-model comparisons. To circumvent these problems, Figure 3 excludes the BMI criterion, and graphically depicts change-in-adjusted-R-squared. In Figure 3 each dot (data-point) represents one of the 231 specific-inventory on-specific-outcome R-squared change values.

A far bigger problem is that contributions to prediction of numbers of variables, items, and words are unaccounted for. The relation of the number of variables to predictive capacity depends on how one computes the latter. Table 8 shows the correlations among the R-squared-change values (by either method) and the number of items and words as well as variables, across the 231 pairings of criterion variables and predictor-set. It also includes the ratio of items to variables.

If one relies on the conventional shrunken (adjusted) R-squared values, one sees a slight increase in prediction as one increases the number of predictor variables; .24 correlation between prediction and number of variables ($p < .001$). On the other hand, if one relies on 5-fold cross-validation, this modest-sized correlation vanishes, turning into a nonsignificant -.04 correlation. The two ways of correcting simple R-squared values yield a similar

Table 6. Cross-Validated (and Adjusted) R-squared Values for Health-Related Criteria: Models Developed Within Present Data

Inventory	Risk-avoidant HP	Good HP	Health Concerns	Peak Tobacco Frequency	Peak Alcohol Frequency	Body-Mass Index (BMI)
Big Five Aspect Scales	.13 (.16) 399	.05 (.07) 399	.08 (.11) 399	.01 (.03) 411	.00 (.02) 410	.01 (.02) 246
NEO Facets (Yarkoni)	.16 (.23) 369	.15 (.21) 369	.05 (.13) 369	.03 (.07) 408	.09 (.15) 407	.16 (.27) 228
AB5C-IPIP	.13 (.23) 398	.03 (.12) 398	.06 (.16) 398	.01 (.05) 449	.00 (.07) 448	.00 (.00) 244
IIDL	.21 (.28) 639	.10 (.18) 639	.06 (.14) 639	.00 (.00) 516	.03 (.10) 516	.29 (.42) 341
IPQ, 65 clusters	.21 (.33) 368	.11 (.23) 368	.03 (.16) 368	.03 (.09) 407	.06 (.16) 406	.05 (.20) 228
IPIP 20 Lexical Factors	.14 (.19) 398	.10 (.15) 398	.05 (.10) 398	.02 (.05) 410	.05 (.09) 409	.04 (.08) 246
Adjective-Scales, 23 Factors	.22 (.25) 639	.13 (.17) 639	.07 (.12) 639	.00 (.01) 516	.04 (.08) 516	.43 (.48) 341
Adjective-Scales, 28 Factors	.20 (.24) 639	.10 (.15) 639	.06 (.11) 639	.01 (.03) 516	.05 (.10) 516	.44 (.49) 341
23 Factors Factor Scores	.27 (.30) 639	.16 (.21) 639	.08 (.13) 639	.03 (.06) 516	.08 (.13) 516	.45 (.50) 341
28 Factors Factor Scores	.27 (.31) 639	.16 (.21) 639	.07 (.13) 639	.01 (.04) 516	.10 (.15) 516	.41 (.48) 341

Note. 5-fold cross-validated change in R^2 (after accounting for age and gender) with conventional adjusted R^2 change in parentheses, N for analyses below

pattern of results overall (correlating .95). Nonetheless, in these data with this level of sample size, clearly *k*-fold cross-validation imposes a more severe correction for inflation when there are very many predictors. By this cross-validation approach, having more variables gave no systematic advantage in prediction.

Number of variables is not an index of predictive efficiency. For example, the MPQ, NEO-PI-R, and IPQ have a similar number of items (240 to 276) and thus similar length, but they differ widely in the number of variables (12 vs. 30 vs. 65). To gauge efficiency, it is better to compare R-squared-change values with the number of items and the total number of words in the inventory. The number of items showed a small positive correlation with predictive capacity, but this was only statistically significant ($p < .05$) with the adjusted-R-square change values (r = .15). There was no meaningful relation between the number of words and predictive capacity. However, as long as one relied on the 5-fold cross-validation coefficients rather than conventional shrunken R-squared, there was a small predictive benefit from aggregation of items (higher item:variable ratio, the MPQ having the highest ratio). Inspection of bivariate scatterplots gave no indication of nonlinear relations among these variables.

One can aggregate the prediction coefficients for each inventory into a single mean across the criteria and correlate across these; this is analogous to aggregating 231 individual scores into means for 21 distinct groups the individuals belong to, and then correlating variable across the 21 group means. By doing so, the size of these associations increases, but there are still only two significant associations across the 21 inventories. Number of variables predicts averaged adjusted-R-squared values (r = .63), which may reflect overfitting. Items-per-variable predicts averaged 5-fold cross-validation values (r =.53). Items-per-variable and number of variables are negatively but non-significantly correlated with each other (r = -.43), which suggests that inventories tended to be of relatively similar length, but differing in how many scales items were being aggregated into. And across 21 inventory-means, the averages of adjusted R-squared and cross-validated R-squared are positively and significantly correlated (r =.63). As Table 8 makes obvious, it matters a great deal how one corrects R-squared values, at least at the moderate sample sizes analyzed here. Because of the moderate sample size here, the cross-validational estimates may well underestimate the R-squared that might arise in new data. Researchers have mainly attributed this phenomenon to a greater instability in estimates when already modest sample sizes are subdivided (Braga-Neto & Dougherty, 2004; Varoquaux, 2018).

Considering predictive efficiency, an important question arises. Conceptually, is it better to represent item cost as number of items or number of words? The number-of-items index assumes that participants take as long to read and

DOI: 10.1002/per

Table 7. Predictive Coefficients (Adjusted R-squared) for Psychopathology-Related Criteria: Models Developed Within Present Data

Inventory	Life-Satisfaction		CES Depression		Mental Health		Fear Questionnaire		Somatoform Dissociation		Self-Report Psychopathy	
Big Five Aspect Scales	.18	(.21) 440	.22	(.25) 439	.13	(.15) 439	.13	(.15) 411	.14	(.17) 435	.02	(.05) 403
NEO Facets (Yarkoni)	.21	(.28) 406	.19	(.25) 405	.11	(.18) 407	.11	(.18) 408	.14	(.20) 404	.02	(.09) 400
AB5C-IPIP	.15	(.24) 435	.12	(.22) 435	.07	(.15) 447	.04	(.12) 449	.11	(.18) 432	.01	(.09) 441
IIDL	.20	(.31) 557	.14	(.24) 617	.09	(.18) 568	.08	(.17) 516	.12	(.21) 582	.00	(.08) 508
IPQ, 65 clusters	.19	(.32) 405	.19	(.30) 404	.14	(.26) 406	.13	(.24) 407	.16	(.27) 402	.07	(.20) 399
IPIP 20 Lexical Factors	.23	(.28) 439	.19	(.23) 439	.09	(.14) 438	.15	(.19) 410	.13	(.17) 434	.06	(.11) 402
Adjective-Scales - 23 Factors	.24	(.29) 557	.20	(.24) 617	.08	(.12) 568	.10	(.15) 516	.16	(.20) 582	.02	(.06) 508
Adjective-Scales - 28 Factors	.23	(.29) 557	.22	(.27) 617	.09	(.14) 568	.10	(.15) 516	.14	(.20) 582	.02	(.07) 508
23 Factors Factor Scores	.27	(.32) 557	.21	(.26) 617	.08	(.12) 568	.11	(.16) 516	.18	(.23) 582	.04	(.09) 508
28 Factors Factor Scores	.25	(.31) 557	.20	(.26) 617	.09	(.14) 568	.11	(.16) 516	.16	(.23) 582	.04	(.09) 508

Note. 5-fold cross-validated change in R^2 (after accounting for age and gender) with conventional adjusted R^2 change in parentheses, N for analyses below

process a sentence as they do for a single adjective. This may be reasonable if he sentences are very short (i.e., phrases, as in IPIP). The number-of-words index penalizes long, verbose or complex items, but assumes that participants spend as much time processing articles and prepositions as they do single adjectives. The number-of-items index seems better, but the other index adds some additionally relevant information helping to differentiate simple from complexly worded items. So both have some relevance.

Figure 4 depicts the relation of inventory length (in terms of number of items) with predictive capacity, relying on the more conservative 5-fold cross-validation mean (as in Table 8). If predictive capacity per item, averaged across the criteria (excluding BMI), is the best index, the champion here would appear to be the Lex-20-IPIP measure, achieving with only 80 items a change in R-squared of.15 (adjusted R-squared) or.11 (5-fold cross-validation). If predictive capacity per word (again, averaged across criteria) is the best index, the champion would be one of the adjective measures (of 23 or 28 factors) deriving from Preliminary Study 1, which achieved similar (.15 and.11, respectively) values with only 112 to 115 words. The champion in terms of sheer (averaged) predictive capacity was the MPQ (with 5-fold cross-validation as the standard, average.15) or the IPQ (with adjusted R-squared as the standard, averaging.23), but these inventories were much longer: 276 items and 3374 words for the MPQ, 258 items and 1571 words for the IPQ. The IPQ and MPQ needed well over twice as many items and three times as many words to achieve barely a 50% increase at most in predictive capacity over these lexically-oriented inventories. Even if one focuses only on models whose items were selected outside the current data-set, the 20-factor adjective inventory with 80 items was more efficient than longer competitors; a doubling of the length of the inventory never doubled the size of the cross-validated R-square. Short measures, at least where they are based on a set of divergent constructs emphasized in the lexicon, emerge as a faster route to a given level of predictive capacity. In Figure 4, increasing the number of items beyond 100 or so did not reveal dramatic increases in prediction proportional to the increase in length. Though not superior in absolute predictive capacity, the lexicon-derived inventories gave the most predictive value for the time and print-space required to administer them. Though the MPQ had the best absolute predictive capacity, the VIA platform reached nearly the same level with fewer items and nearly half the number of words, making it arguably more efficient. As for the weakest performers, examination of tables 4 to 7 would reveal that the most consistently low prediction values were associated with the 6FPQ inventory. Moreover, Figure 4 makes clear that the AB5C-IPIP was the most inefficient for prediction, unsurprising since its circumplex approach maps all possible varying combinations of Big Five factors rather than going beyond the Big Five.

DOI: 10.1002/per

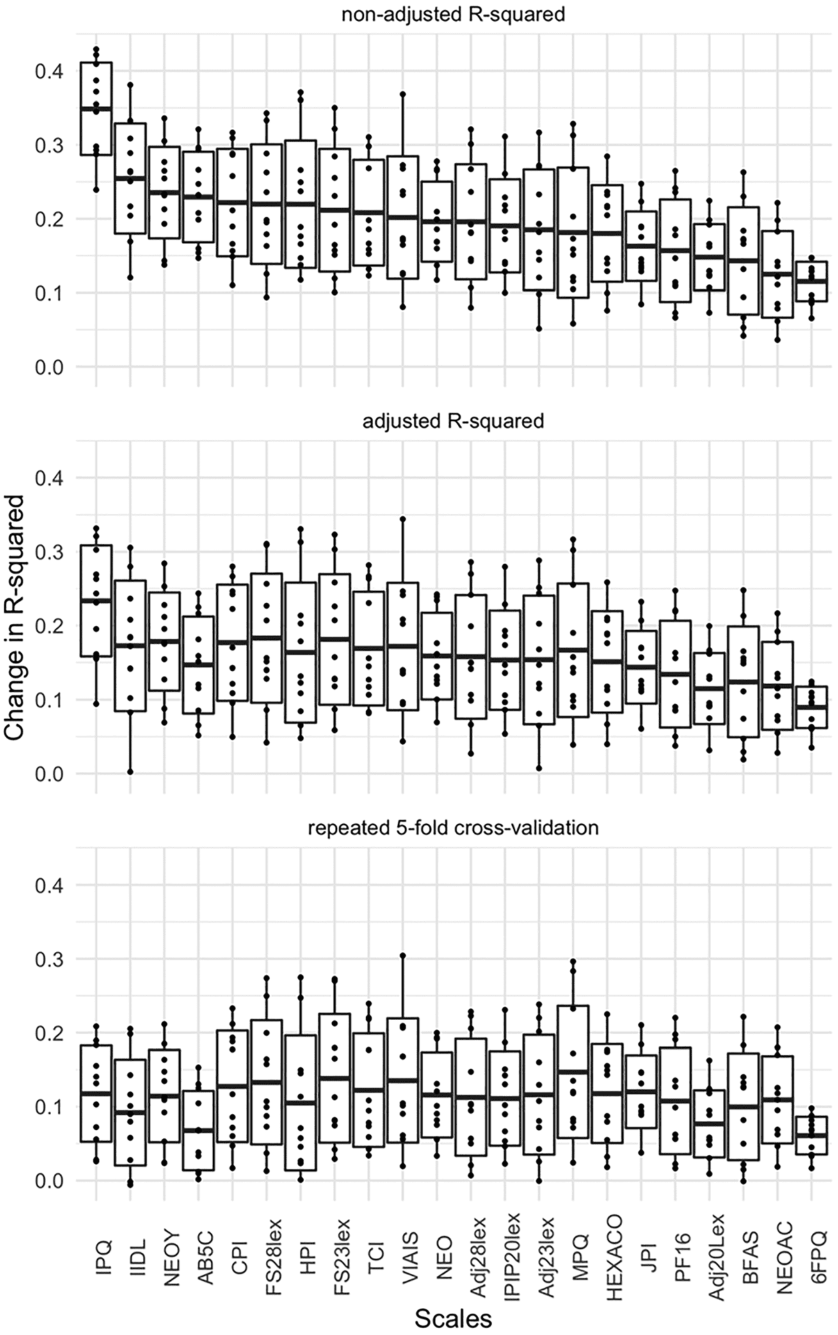

Figure 3. R-squared Change Prediction Coefficients for 22 Inventories, Showing Mean Coefficient and Ordered By Unadjusted R-squared Change

Table 8. Correlations Between Number of Variables, Items, and Words in Inventory and Predictive Capacity for the Inventory

	Change in R^2 5-Fold C.V.	Change in Adj. R^2	No. of Variables	No. of Items	No. of Words	Items Per Variable	Words Per Variable	Words Per Item
R^2 change, 5-Fold C.V.	1.00	0.63**	-0.15	0.35	0.28	0.53*	0.35	0.13
R^2 change, Adjusted R^2	0.95**	1.00	0.63*	0.41	0.11	0.07	-0.10	-0.15
Number of Variables	-0.04	0.24**	1.00	0.18	-0.05	-0.43	-0.40	-0.25
Number of Items	0.10	0.15	0.25**	1.00	0.44*	0.69*	0.22	0.03
Number of Words	0.09	0.04	0.02	0.45**	1.00	0.44*	0.81**	0.83**
Items Per Variable	0.16*	0.03	-0.40**	0.69**	0.43**	1.00	0.64*	0.24
Words Per Variable	0.11	-0.04	-0.38**	0.24**	0.81**	0.64**	1.00	0.81**
Words Per Item	0.04	-0.06	-0.19**	0.06	0.83**	0.25**	0.81**	1.00

Note. Values below the diagonal are across 11 criteria (excluding BMI) for each of 21 inventories (N=231). Values above the diagonal involve only the inventory's mean across 11 criteria, for each of 21 inventories (N=21). * $p < .05$, ** $p < .01$

DOI: 10.1002/per

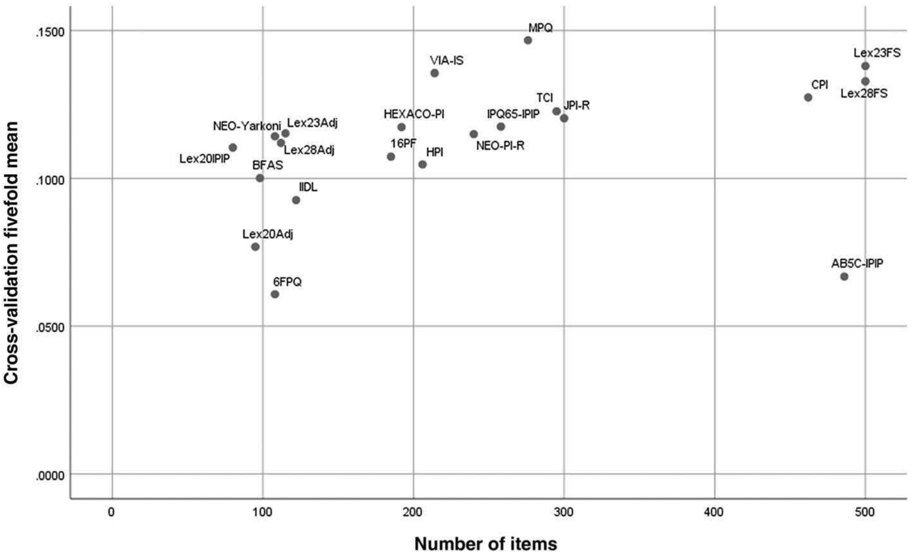

Figure 4. Scatterplot Depicting Relation of Number of Items in an Inventory with Cross-Validated Predictive Capacity

Worth acknowledging is the possibility that performance in these comparative-validity tests was affected by the nature of the criterion variables, which emphasized health and mental health. It could be that the MPQ, in particular, was built to be more responsive to these domains than are inventories targeted more to industrial-organizational settings (e.g., the HPI) or reflecting the imprint of a classification of psychological needs (e.g., the 6FPQ).

Moreover, we used the maximum sample size available for each combination of variables. Had we run all analyses on that small subsample that had values on all variables, estimates might differ somewhat, but due to decreased statistical power would tend to be less generalizable. In designing future studies of this type, it would be desirable to have a larger sample and equalize sample sizes for all analyses.

Discussion

The main study in this report is constructed as a kind of "horse race" among competing, alternative assessment platforms. One way to report on a horse race, of course, is to report who were the winners and who were the losers. A deeper, and in the long run more pertinent accounting would consider the implications of the particular horse-race outcome for the broader context of optimal horse culture, breeding, and racing, and what kind of organization of such activities is best. Analogously, we could focus entirely on which instruments appeared to obtain stronger validation than others, but instead we put more emphasis on what results suggest about how to construct predictively powerful, efficient personality measures, and what kind of structure (or organization) of personality variation provides the best foundation.

The models compared involved at least five distinct scale-construction philosophies. Under an empirical criterion-keying approach (e.g., the CPI), scales are typically large aggregations of items such that each item contributes to the prediction of some external criterion. Under a domain-and-facet approach (e.g., the NEO-PI-R), the facet scales are selected as sets of differentiated indicators related to one overarching domain. The MPQ was constructed by sequentially constructing a set of facet-level scales without emphasizing any particular *a* priori broad domains, only inferring domains later (Tellegen & Waller, 2008); essentially this represents a facets-without-domains approach. Under a high-dimensionality (HD) approach (e.g., the Lex-20), scales reflect a large number of relatively independent dimensions that are relatively robust, allowed to be more independent than facets might be, and with no separation between domains and facets. A cluster-based approach tends to resemble the HD approach, except that cluster analysis employed without any robustness standard tends to allow derivation of a larger number of granular constructs than do factor-analytic approaches.

There was little distinct support for the criterion-keying approach; although the CPI demonstrated decent predictive capacity, this came at the cost of a comparatively large number of items and by far the largest number of words, indicating relatively poor return on investment of time and print-space. Inventories built on a domain-and-facet approach did not stand out as superbly predictive, and two of these (AB5C and 6FPQ) were the worst performers. The facets-without-domains approach had better support, given the good performance of the MPQ. The high-dimensionality factor-analytic approach – seeking many relatively independent factors or variables without consolidating them into broad domains – included the inventories tending to have the best predictive capacity especially in terms of efficiency. For adjective measures, the wider-variable-selection set of 23 and 28 factors predicted better than the set of 20 narrow-selection factors (though not much better than the IPIP scales constructed to capture those 20 variables). The cluster-based approaches showed superior performance if one simply examined the uncorrected

DOI: 10.1002/per

R-square values, or used a shrunken R-squared index, but less outstanding performance with 5-fold cross-validation to correct for overfitting. The evaluation of this approach (and the IIDL and IPQ which drew on it) seems to depend on how one comparatively evaluates two ways of correcting for overfitting. Moreover, both the IIDL and IPQ were originally derived in the same data in which they were tested here, which only emphasizes further that they cannot be recommended firmly without scrutiny in further data, beyond the scope of this report.

We cannot conclude, based on these results, that having more words, or more items, or even necessarily more variables, yields systematic improvement in prediction. Comparatively strong results can be obtained by seeking a dozen or two well-selected variables without forcing them or pruning them to function as indicators of a few broad domains. It would seem attractive to adopt a model with far more variables than that (e.g., here 61-65 of them), but there may be little improvement in prediction if rigorous corrections for overfitting are applied. Possibly, such very-high-predictor-number models will show their merits more clearly in extremely large samples, where hold-out validation samples can be quite large. Moreover, their merits may be clearer when cross-validation methods use a much higher *k* value, another issue that needs further exploration.

Generally, results tend to be consistent with a dictum attributed to Frederic Lord, that one maximizes (predictive) validity by maximizing the number of predictors while minimizing their intercorrelation, even though one maximizes reliability by an opposite tack (maximizing the intercorrelation of items). This dictum particularly helps account for predictive efficiency. However, maximizing absolute predictive capacity may gain some benefit from internal-consistency reliability (a higher item:variable ratio), if only because this kind of reliability yields better cross-validation of associations with criteria at moderate sample sizes. Still, the prediction gains made along that avenue are not very efficiently gained. One sacrifices large quantities of participant time, and space in questionnaires, to achieve them. The best-performing instruments here tended to be those derived not by a reliability-maximizing approach, but rather those that sought many predictor variables with minimal intercorrelation. One might say these are highly divergent "facets" unmoored to broad domains, although without the aggregation into domains, it is no longer sensible to call them facets of anything.

As noted earlier, we limited our predictor models to those having multiple items aggregated to measure each variable, because of advantages in the direction of interpretability, validity, and potentially theory. Comparisons with single-item prediction were thus outside our scope, but it would be useful to make a systematic comparison of single-item-variable versus multiple-item-variable prediction along the same lines as here, while taking account of the advantages and disadvantages (beyond sheer predictive capacity) of these respective approaches. Our results do imply that rigorous cross-validation to correct for overfitting might wipe out much of the apparent predictive advantage single-item-variables could offer. But this is only our estimate based on our study of a single community sample of moderate sample size. More data should be examined, especially data with much larger sample sizes than employed here. It may well be that in the case of huge sample sizes single-item-prediction becomes superior. Also beneficial would be examination of cross-validation with more than 10 folds, in case this would lessen somewhat the discrepancy between the adjusted R-squared values (which here tended to favor inventories with more variables) and cross-validated values (which here tended to favor longer scales). This discrepancy is an interesting and fairly novel finding here, that deserves further future exploration.

CONCLUSIONS

In personality science, there has been a tendency to favor parsimonious structures with a few broad domains or dimensions, and rely on inventories that are relatively long, with many items and words. But as our preliminary study indicated, the structure of personality appears more complex than has been previously assumed. And inventories with a dozen or two (or perhaps more) highly divergent variables demonstrably improve prediction. So parsimony tends to compromise prediction.

Recommendations that would emerge from this study take one beyond heavy emphasis on parsimony. Building a personality model or its associated measurement instrument by first identifying a few broad domains, then subdividing those (perhaps symmetrically into three or four or six subcomponents per domains) may serve other goals, but this approach does not appear to optimize predictive capacity. For optimizing that, it may be better to build out whatever number of variables seems sufficiently comprehensive of the field of relevant and important variables, and let any broader domains emerge *a posteriori* rather than *a priori*. For questionnaire assessment of relevant variables, in absolute terms more items may better the prediction, but merely maximizing item number compromises efficiency. Efficiency is better served by keeping number of items moderate (e.g., approximately 100 may be enough), ensuring that they represent (at least) a *large* number of dimensions of personality variation evident in natural-language factors, and selecting these items based on convergent and discriminant properties with respect to these factors. Thus, we suggest that prediction-favoring personality models will simply look different than the kind of models that became conventional over the course of the first century of personality science.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1 Supporting Information

Supporting info item

DOI: 10.1002/per

REFERENCES

Afendras, G., & Markatou, M. (2019). Optimality of training/test size and resampling effectiveness in cross-validation. *Journal of Statistical Planning and Inference*, *199*, 286–301.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycholexical study. *Psychological Monographs*, *47*, (Whole No, 211.

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Toward a taxonomy of personality descriptors in German: A psycho-lexical study. *European Journal of Personality*, *4*, 89–118.

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior*, *19*, 289–303.

Benet-Martinez, V., & Waller, N. G. (2002). From adorable to worthless: Implicit and self-report structure of highly evaluative personality descriptors. *European Journal of Personality*, *16*, 1–41.

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*, 374–380.

Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., & Wetzel, R. D. (1994). *The Temperament and Character Inventory (TCI): A guide to its development and use*. St. Louis, MO: Center for the Psychobiology of Personality, Washington University.

Conn, S. R., & Riek, M. L. (1994). *The 16PF fifth edition technical manual*. Champagin, IL: Institute for Personality and Ability Testing.

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

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper/Collins.

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*, 880–896.

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*, 84–96.

Goldberg, L. R., & Saucier, G. (2016, January). The Eugene-Springfield Community Sample: Information available from the research participants. *ORI Technical Report*, *v. 56*(no. 1).

Gough, H. G. (1996). *CPI manual* (Third ed.). Palo Alto, CA: Consulting Psychologists Press.

Hofstee, W. K. B., De Raad, B., & Goldberg, L. R. (1992). Integration of the Big-Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*, 146–163.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

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

Jackson, D. N. (1994). *Jackson Personality Inventory – revised manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire manual*. Port Huron, MI: Sigma Assessment Systems.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, *39*, 329–358.

Marks, I. M., & Mathews, A. M. (1979). Brief standard self-rating for phobic patients. *Behavior Research and Therapy*, *17*, 263–267.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H., Wainer & H. I., Braun (1988), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Nijenhuis, E. R. S., Spinhoven, P., van Dyck, R., van der Hart, O., & Vanderlinden, J. (1997). The development of the Somatoform Dissociation Questionnaire (SDQ-5) as a screening instrument for dissociative disorders. *Acta Psychiatrica Scandinavica*, *96*, 311–318.

Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population*. Department of Psychology, University of Michigan.

Paulhus, D. R., Neumann, C. S., & Hare, R. (in press). *Manual for the Self-Report Psychopathy Scale (SRP-III)*. Toronto: Multi-Health Systems.

Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the Big Five? Plenty. *Journal of Personality*, *68*, 821–835.

Pavot, W., & Diener, E. (1993). Review of the Satisfaction With Life scale. *Psychological Assessment*, *5*, 164–172.

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. New York: Oxford University Press/Washington. DC: American Psychological Association.

Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measuremnet*, *1*, 385–401.

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, *73*, 1296–1312.

Saucier, G., & Goldberg, L. R. (1998). What is beyond the Big Five? *Journal of Personality*, *66*, 495–524.

Saucier, G., & Iurino, K. (2019, in press). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology*.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. *The SAGE handbook of personality theory and assessment*, *2*, 261–292.

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*(pt. A), 68–77.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. *Journal of Research in Personality*, *44*, 258–272.

Yarkoni, T. (2009). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*, 180–198.

DOI: 10.1002/per