Psychological Assessment

Using a Genetic Algorithm to Abbreviate the Psychopathic Personality Inventory–Revised (PPI-R)

Hedwig Eisenbarth, Scott O. Lilienfeld, and Tal Yarkoni

Online First Publication, December 1, 2014. http://dx.doi.org/10.1037/pas0000032

Psychological Assessment

1040-3590/14/$12.00 http://dx.doi.org/10.1037/pas0000032

Using a Genetic Algorithm to Abbreviate the Psychopathic Personality Inventory–Revised (PPI-R)

Hedwig Eisenbarth
University of Colorado Boulder
and University of Regensburg

Scott O. Lilienfeld
Emory University

Tal Yarkoni
University of Texas at Austin

Some self-report measures of personality and personality disorders, including the widely used Psychopathic Personality Inventory–Revised (PPI-R), are lengthy and time-intensive. In recent work, we introduced an automated genetic algorithm (GA)-based method for abbreviating psychometric measures. In Study 1, we used this approach to generate a short (40-item) version of the PPI-R using 3 large-*N* German student samples (total $N = 1{,}590$). The abbreviated measure displayed high convergent correlations with the original PPI-R, and outperformed an alternative measure constructed using a conventional approach. Study 2 tested the convergent and discriminant validity of this short version in a fourth student sample ($N = 206$) using sensation-seeking and sensitivity to reward and punishment scales, again demonstrating similar convergent and discriminant validity for the PPI-R-40 compared with the full version. In a fifth community sample of North American participants acquired using Amazon Mechanical Turk, the PPI-R-40 showed similarly high convergent correlations, demonstrating stability across language, culture, and data-collection method. Taken together, these studies suggest that the GA approach is a viable method for abbreviating measures of psychopathy, and perhaps personality measures in general.

Keywords: psychopathy, genetic algorithm, abbreviation, personality

Personality assessment using self-report questionnaires can be a time-consuming process. To maximize reliability and validity, the developers of personality inventories often use a large number of items to assess each construct, in some cases producing decidedly unwieldy instruments when multiple constructs are targeted. The assessment of psychopathic personality traits is no exception. Although a few self-report measures of psychopathy, such as the Levenson Primary and Secondary Psychopathy Scales (Levenson, Kiehl, & Fitzpatrick, 1995), are relatively brief (26 items), most are considerably lengthier. For example, the Triarchic Psychopathy Measure (Patrick, Fowles, & Krueger, 2009) contains 58 items, and perhaps the most widely used self-report measure of this construct, the Psychopathic Personality Inventory–Revised (PPI-R), contains 154 items (PPI-R; Lilienfeld & Widows, 2005). Logistical considerations can hamper researchers' ability to administer a measure such as the PPI-R; participants' time is often in short supply, and any time spent filling out the PPI-R cannot be spent completing other questionnaires or tasks. Consequently, there are good reasons to develop abbreviated versions of psychopathy measures that can be more easily applied in settings where resources are limited, or in large-scale epidemiological work where questionnaire space is often at a premium.

The desire to develop short versions of existing measures is offset by the need to measure constructs' high fidelity. As Cronbach (1954) noted, given a fixed number of items, there is typically a tradeoff between bandwidth and fidelity. Other things being equal, the act of substantially shortening a measure is likely to reduce its reliability and/or validity. Consequently, there is a need for methods that can optimize the balance between brevity and fidelity in a quantitative, systematic, and efficient way. In recent work, Yarkoni (2010) introduced a novel, highly efficient, and almost completely automated approach to the abbreviation of questionnaire measures. Yarkoni used a *genetic algorithm* (GA)—a programmatic approach that uses evolutionary princi-

Hedwig Eisenbarth, Department of Psychology and Neuroscience, University of Colorado Boulder and Department of Forensic Psychiatry and Psychotherapy, University of Regensburg; Scott O. Lilienfeld, Department of Psychology, Emory University; Tal Yarkoni, Department of Psychology, University of Texas at Austin.

Hedwig Eisenbarth is now solely at the Department of Psychology and Neuroscience, University of Colorado Boulder.

Dr. Hedwig Eisenbarth is coauthor of the German version of the PPI-R and receives royalties for the sale of this measure, which is published by the Hogrefe Verlagsgruppe in Goettingen, Germany. Dr. Scott O. Lilienfeld is the codeveloper of the Psychopathic Personality Inventory–Revised, and receives royalties for the sale of this measure, which is published by Psychological Assessment Resources in Lutz, Florida. This research was supported in part by a fellowship for Hedwig Eisenbarth from the German Research Foundation (DFG).

Correspondence concerning this article should be addressed to Hedwig Eisenbarth, Department of Psychology and Neuroscience, University of Colorado Boulder, 344 UCB, Boulder, CO 80302. E-mail: hedwig.eisenbarth@colorado.edu

ples to progressively "evolve" high-quality solutions in contexts where the dimensionality of a problem is too high to afford an analytical solution. Yarkoni demonstrated that the technique could be profitably applied to a broad range of personality measures, with particular benefits for long measures. For instance, Yarkoni generated a 181-item instrument that accurately recaptured variance in over 200 different scales with approximately 1,000 items drawn from eight broadband inventories, representing a 90% savings in instrument length. This method is based on the idea of lowering redundancy within a scale, and therefore reducing the items to the substrate that does best in capturing the traits of interest.

In the present work, we used Yarkoni's (2010) GA-based approach to produce an abbreviated version of the PPI-R. The PPI-R is a frequently used questionnaire for assessing psychopathic personality traits; however, it is relatively long, at 154 items, and although there are unpublished short versions, there are no published short forms. There is one recent report on the development of an abbreviated version of the original PPI (which contains 187 items). This short form, called the PPI-SF (Tonnaer, Cima, Sijtsma, Uzieblo, & Lilienfeld, 2012), was produced using Mokken scale analysis (MSA), which is an Item-Response-Theory-based model. The PPI-SF resembles the PPI in its ability to discriminate a forensic population from controls, and shows similar correlations to the PPI with the well-validated Psychopathy Checklist–Revised (PCL-R; Hare, 2003), a semistructured interview that incorporates file data. However, the PPI-SF has at least two remaining limitations. First, it remains relatively long; at 100 items, it is not suitable for many applications requiring more rapid assessment of psychopathic traits. Second, it cannot be readily adapted to produce alternate forms or even shorter versions. An important feature of the GA-based abbreviation approach is that a measure can be abbreviated to varying degrees by varying a single free parameter (representing the relative cost of retaining each additional item; see Yarkoni, 2010 for details). Thus, one can readily generate a range of different measures and select the abbreviated version that a researcher deems optimal in terms of balancing brevity and fidelity.

Here we report three studies in which we generated and validated a novel, very short (40 Items) version of the PPI-R. In Study 1, we applied Yarkoni's (2010) GA approach to generate an abridged version of the PPI-R using three large-*N* German student samples. In Study 2, we tested the convergent and discriminant validity of this short version in a fourth student sample using sensation-seeking and sensitivity to reward and punishment scales. Finally, in Study 3, we used a fifth sample to demonstrate that the abbreviated measure is robust to differences in culture, language, and data-collection method.

Method

Samples

For the initial measure abbreviation (Study 1), data from three samples of college students were included. The first data set was derived from a large survey of 491 students of different majors (age: $M = 23.57$, $SD = 3.61$; 277 females, 214 males). The second sample comprised 721 students of mainly economics majors (age: $M = 22.03$, $SD = 2.09$; 413 females, 308 males). The third sample comprised 481 students and was acquired for validation of a different measure, the TriPM (age: $M = 24.25$, $SD = 2.64$; 272 females, 209 males). The first three samples were used to generate the abbreviated measure, which we term the PPI-R-40 (Study 1). To maximize data integrity and avoid using imputation procedures that might bias the abbreviation process, we excluded all subjects who omitted a response to at least one item on the original PPI ($n = 99$ across all three samples). We additionally excluded one clear outlier who responded "1" on all questions (producing a *z*-score of −7.6 below the mean on the total PPI-R). Thus, the total sample size across all three data sets was 1,590 (911 males, 679 females; including all omitted subjects by applying mean imputation had no discernible impact on the results reported here). It is important to note that the heterogeneity of the combined sample does not present any problem with respect to our abbreviation approach. To the contrary, since our goal was to generate a measure that would generalize well to different populations, diversity in the training sample was a desirable feature.

For the validation studies (Studies 2 and 3), we used data from two additional samples. In Study 2, we used data from a sample of 206 students (age: $M = 23.44$, $SD = 3.51$; 155 females, 51 males), derived from a study on reward sensitivity, to evaluate the convergent and discriminant validity of the PPI-R-40 in relation to other measures relevant to psychopathy (Study 2). No participants had to be excluded from this sample, as the measures were administered online and participants could not skip questions. In Study 3, we used a general population sample recruited broadly across the North American community (United States and Canada) acquired using Mechanical Turk, an online data collection platform hosted by Amazon. Like most M-Turk samples (Paolacci, Chandler, & Ipeirotis, 2010), this was a well-educated sample, with 89% of participants having attended at least some college and 47% having attained at least a bachelor's degree. Sixty-eight percent of the sample was employed, and an additional 16% of the sample were students. We used the original, English-language version of the PPI-R ($N = 239$, age: $M = 33.06$, $SD = 11.16$; 138 females, 101 males) to ensure that the PPI-R-40 was robust across translations, cultures, and data-collection methods (online vs. offline). We again excluded participants with any missing data ($n = 73$ for this sample; as above, retaining all subjects and imputing missing responses did not meaningfully alter results).

Measures

The main measure included in this study was the PPI-R (German version: Alpers & Eisenbarth, 2008; English original: Lilienfeld & Widows, 2005). This self-report questionnaire of psychopathic personality traits consists of 154 items that can be assigned to eight subscales and three validity scales designed to detect aberrant responding. The subscales are Blame Externalization, Rebellious Nonconformity, Coldheartedness, Social Influence, Carefree Nonplanfulness, Fearlessness, Machiavellian Egocentricity, and Stress Immunity. These factor-analysis-derived subscales have been shown to be assignable to two main factors: Fearless Dominance and Self-Centered Impulsivity, sometimes also called Impulsive Antisociality (Benning, Patrick, Hicks, Blonigen, & Krueger, 2003; Neumann, Malterer, & Newman, 2008). In addition, the PPI-R includes Deviant and Virtuous Responding scales, both intending to measure response biases. The German version

has demonstrated good internal consistency of $r_\alpha = .85$ for the total score in students and incarcerated samples (Alpers & Eisenbarth, 2008).

To evaluate convergent and discriminant validity (Study 2), we used data from a questionnaire on sensitivity to punishment and sensitivity to reward, as well as a subscale of a personality questionnaire of sensation seeking. To measure reward and punishment sensitivity, the Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ; Torrubia, Avila, Molto, & Caseras, 2001) was used in its German version (Hewig, Hagemann, & Riemann, 2014). This 48-items questionnaire includes two subscales, one on sensitivity to reward and one on sensitivity to punishment, referring to the behavioral activation and behavioral inhibition systems, respectively. According to some authors (e.g., Lykken, 1995), psychopathy is associated with deficient behavioral inhibition system functioning but intact or perhaps overactive behavioral activation system functioning. The content validity of this questionnaire has been demonstrated in relation to neuroticism, extraversion, and anxiety scales (see Torrubia et al., 2001). As a measure for sensation seeking, the subscale of Sensation-Seeking of the Zuckerman-Kuhlman Personality Questionnaire (Zuckerman, 2002) was administered in its German version (Ostendorf & Angleitner, 1994). The questionnaire consists of five subscales: Neuroticism, Activity, Sociability, Impulsive Sensation-Seeking and Aggression/Hostility. The subscale Impulsive Sensation-Seeking includes 19 items and has been shown to have a positive relationship to Extraversion and a negative relationship to Conscientiousness. This measure was administered in view of research demonstrating positive associations between psychopathy and sensation seeking (e.g., Fulton, Marcus, & Payne, 2010).

Measure Development

The GA-based abbreviation procedure we used is described in detail in Yarkoni (2010); here we review only key elements of the process. We have implemented all procedures in an open-source Python package called *scythe*, freely available on GitHub at http://github.com/tyarkoni/scythe. The package bundles nearly all of the data, analyses, and results reported here into a tutorial presented as an IPython Notebook; thus, users can easily reproduce our work or adapt our code for their own purposes.

We begin by coding each item on a to-be-abbreviated measure as a single bit (or gene) on a "chromosome" containing all 154 items. Each bit can be turned on or off (0 or 1), indicating whether the abbreviated measure should include or exclude the corresponding item. For example, if a target measure has eight items, an abbreviated version of that measure that retains only the second, fifth, and eighth items would be represented as 01001001. A measure that retains the first, second, and third items would have the representation 11100000, and so on. Each of these representations is referred to as an *individual.*

After generating an initial population of, say, 200 random individuals, we use each individual to generate a different abbreviated scoring key. This is accomplished simply by selecting the N items on the abbreviated measure that show the strongest absolute correlation with each target subscale on the full-length measure. For example, suppose an individual chromosome contains 60 ones and 94 zeroes—meaning that 60 items are to be retained from the initial starting pool of 154 PPI-R items. For each of the PPI-R scales, we order all 60 of the retained items by their descending absolute correlation with the original subscale score, and define the abbreviated scoring key for the scale as simply the linear sum of the first N items in the list. For example, if $N = 3$, and items 14, 23, and 36 have the strongest absolute correlations with the Fearlessness scale, with rs of 0.3, 0.22, and -0.21, then the abbreviated scoring key for Fearlessness would be Item 14 + Item 23 – Item 36 (reflecting the fact that the third item is inversely correlated). Thus, we end up with 200 different abbreviated measures in each generation.

Naturally, some of these randomly produced abbreviated measures are bound to be better than others. Since our goal is to evolve an increasingly good abbreviation, only the best individuals in each generation are used to populate the next generation of individuals, subject to some degree of variation through random recombination and mutation. Recombination is achieved by splicing two individuals; mutation is achieved by "flipping" bits randomly (e.g., the individual 11110000 would become 11110001 if mutation randomly affected the last bit). The selection and variation process is iterated for some fixed number of generations, or until the fitness of the best individual asymptotes, at which point the single fittest individual in the last generation is taken to represent the solution to the problem—that is, the final abbreviated measure.

The fitness of each individual within each generation is assessed via a predefined loss function. The loss function used in Yarkoni (2010) and in the present work can be thought of as the sum of two quantities: (a) an *item cost*, which increases in direct proportion to the number of items retained by the abbreviated measure, and (b) a *variance cost*, which increases in proportion to the amount of unexplained variance in the original, full-length, measure. Because these two quantities are in direct tension (i.e., shorter measures will necessarily recapture less variance in the original measure), minimizing the overall loss function requires optimization of the balance between brevity and fidelity. Formally, the loss function can be expressed as:

$$Loss = Ik + \sum_{i=1}^{s} 1 - R_i^2 \quad (1)$$

Where I is a free parameter Item Cost (IC) determined by the investigator, k is the number of items retained by the GA, s is the number of scales in the inventory, and R_i^2 is the amount of variance in the ith scale that can be explained by a linear, unit-weighted sum of individual item scores. Crucially, by varying the Item Cost parameter, one can place a greater or lesser emphasis on the brevity of the measure relative to its fidelity. When I is high, the cost of each additional item outweighs the cost of a loss in explained variance, leading to a relatively brief measure. Conversely, when I is relatively low, the GA has little incentive to remove items, leading to a longer measure that maximizes explained variance.

In the present work, we systematically varied the cost parameter to illustrate the flexibility of the method and the impact of this choice on the results. The GA was instructed to recapture variance in the eight subscales of the PPI-R; the Virtuous and Deviant Responding scales of the PPI-R were excluded from the analyses, as these scales are not content subscales, but are instead intended to test for potentially invalid response tendencies (i.e., socially

desirable responding, random or careless responding). Therefore the GA was based only on the eight content scales.

To ensure that our assessment of the quality of any abbreviated measure was unbiased, and that the resulting measure generalized across multiple samples, we used a cross-validation approach. Specifically, we randomly divided the pooled data from all three samples in Study 1 into training and testing halves. We used the training half to generate the measure, and then applied that measure to the testing half in order to quantify performance. Note that this approach, although statistically unbiased, underestimates the true validity of the measure, as it does not use all of the available information in the construction process. To maximize fidelity to the original measures, the actual item list we report for the final abbreviated measure was based on the full collapsed sample without cross-validation, ensuring that the GA took advantage of all of the available data (this two-pronged approach—of quantifying model performance in a cross-validated way, but fitting the final model using all available data—is standard in the machine learning literature). In addition, the validity of the PPI-R-40 was tested in an independent sample (Study 2), correlating the scores and comparing the reliability coefficients, as well as correlating the short version subscales scores with an external measure, the SPSRQ. Another validity test (Study 3) was conducted, testing the version in an English-speaking sample with the English version of the PPI-R.

Results

Study 1

Our initial analysis sought to establish the basic efficacy and generalizability of our abbreviation approach. We used all available data (n = 1,590 across three samples) to develop an abbreviated version of the PPI that successfully recaptured most of the variance in the original measure. To illustrate the flexibility of our approach and examine the trade-off between measure brevity and measurement fidelity, we systematically varied two parameters. First, we set the item cost (IC) parameter, which controls the degree to which the abbreviation process emphasizes brevity over fidelity, to 0.02, 0.04, 0.06, or 0.08. Second, we manipulated the maximum number of items (MI) that could be used to score each scale, using values of 3, 5 7, or 9. The combination of these two parameters resulted in 16 different configurations of the genetic algorithm. (Note that these parameter ranges were chosen so as to produce a reasonably broad range of solutions. Given that the need for brevity vs. fidelity varies across contexts, there is no principled way to select a single optimal combination of settings.) In each configuration, we allowed the genetic algorithm to evolve a solution over 1,000 generations.

Table 1 displays key properties of the final measure produced in each configuration. As expected, varying the IC exerted a robust effect on the brevity/fidelity tradeoff, with higher values producing longer measures that retained more of the variance of the original PPI-R. Varying the number of items allowed to load on each scale had less predictable effects, with low values generally producing better results at low ICs and high values producing better values at high ICs. Importantly, fidelity was high even with relatively short measures. For example, even an instrument with just 22 items (IC = .08, MI = 9) produced a mean convergent correlation of .65 across the eight original PPI-R scales. Note that our abbreviation approach allows individual items to be used in scoring multiple subscales, thereby potentially producing measures with fewer items than one might expect (e.g., a measure with MI = 7 could have fewer than 7 × 8 = 56 items in total; see Table 1). However, because the PPI-R-40 measure that we ultimately deemed optimal did not reuse any items, we do not discuss this point further (for further discussion, see Yarkoni, 2010).

Table 1
Genetic Algorithm Parameters and Resulting Item Selections

MI	IC	No. of items	Mean R^2	Mean alpha
3	0.02	24	0.72	0.61
3	0.04	24	0.71	0.61
3	0.06	23	0.7	0.6
3	0.08	22	0.68	0.57
5	0.02	40	0.82	0.69
5	0.04	39	0.8	0.68
5	0.06	34	0.78	0.66
5	0.08	31	0.75	0.64
7	0.02	54	0.86	0.73
7	0.04	48	0.84	0.72
7	0.06	44	0.82	0.69
7	0.08	31	0.72	0.63
9	0.02	67	0.9	0.76
9	0.04	55	0.86	0.75
9	0.06	42	0.77	0.71
9	0.08	22	0.65	0.56

Note. IC = Item cost; MI = Maximum number of items per subscale.

Our motivation in systematically varying these parameters was to illustrate the ease with which researchers can tailor an abbreviated measure to their specific goals. However, because we deemed it imprudent to release 16 different abbreviated versions of the PPI "into the wild," we opted to focus on one configuration (IC = 0.02, MI = 5) that in our view effectively balanced brevity and fidelity, with a total of 40 retained items and a mean convergent correlation of r = .91 (range = .85 to .93, see Table 2) with the parent measure. We term this measure the PPI-R-40.

Several additional analyses attested to the validity of the PPI-R-40. First, as evident in the pattern of intercorrelations among scales, this measure retained most of the specificity of the original measure (see Figure 1). Second, visual inspection of scatterplots revealed strong linear relations between original and abbreviated scores for all scales (see Figure 2). The mean internal consistency (Cronbach's alpha) for the eight PPI-R subscales was .69 (range = .55 to .74); for the factor scale Fearless Dominance and for the factor scale Self-Centered Impulsivity, the alphas were .78 and .72, respectively. Given that the abbreviation process is explicitly designed to eliminate redundant variance from a measure, low internal consistency values are a desirable property in the present context—and indeed, when coupled with high convergent correlations, are a sign that the genetic algorithm is operating effectively (for further discussion, see Yarkoni, 2010). Lastly, we fit a polytomous IRT model to both the original and abbreviated measures (using the "graded" model implemented in the *mirt* R package). As Figure 3 illustrates, test information was not meaningfully lower for the abbreviated measure than for the full-length measure at any point in the latent ability curve when accounting for the differing number of items (see dashed gray lines).

Table 2
Convergent Correlations of PPI-R Subscales, Factor Scores, and Sum Score of the Original and the Abbreviated Version (In Brackets Item Numbers From the Original Version) and Cronbach's Alpha for the Abbreviated and the Original Version Using the Three Samples of Study 1 (N = 1,590)

	PPI-R-40 convergent correlations	PPI-R-40 Cronbach's alpha	Original version Cronbach's alpha
Blame Externalization (18,19,40,84,122)	.91	.70	.87
Carefree Nonplanfulness (89,108,121,130,145)	.87	.68	.81
Coldheartedness (27,75,97,109,153)	.92	.72	.85
Fearlessness (12,47,115,137,148)	.93	.74	.86
Machiavellian Egocentricity (33,67,77,136,154)	.85	.55	.79
Rebellious Nonconformity (4,36,58,80,149)	.90	.68	.80
Social Influence (22,34,46,87,113)	.90	.72	.87
Stress Immunity (10,32,76,119,140)	.93	.70	.85
Self-Centered Impulsivity	.92	.71	.88
Fearless Dominance	.95	.78	.90
Sum Score	.95	.79	.91

Cross-validation. The abbreviated measure displayed in Table 2 was generated using all available data, thus minimizing the expected variance of the result. However, the convergent correlation estimates reported earlier (Table 2; Figures 1–2) were potentially susceptible to some degree of overfitting, as the same data were used to generate the measure and evaluate its performance. To provide unbiased estimates of the measure's fidelity, we repeated the generation process using a cross-validated approach by randomly dividing the pooled data from all three samples into training and testing halves. We repeated the abbreviation process on only the training half, and then used the resulting measure and scoring key to assess performance in only the testing half, providing an unbiased estimate of how well the abbreviated measure would generalize to a new set of subjects. It is important that performance decreased only slightly, with a mean convergent R^2 of .82 (range = .73 to .86) for the eight PPI scales (note that this decrease may have reflected the reduction in data available for training rather than overfitting per se).

Comparison with a "top-*N*" heuristic. The abbreviated measure we opted for contained 40 items. Given that this was identical to the theoretical maximum of 40 items (i.e., five different items for each of the eight scales), one might question the utility of GA-based abbreviation compared with the simpler heuristic of retaining the five items that showed the highest zero-order corre-

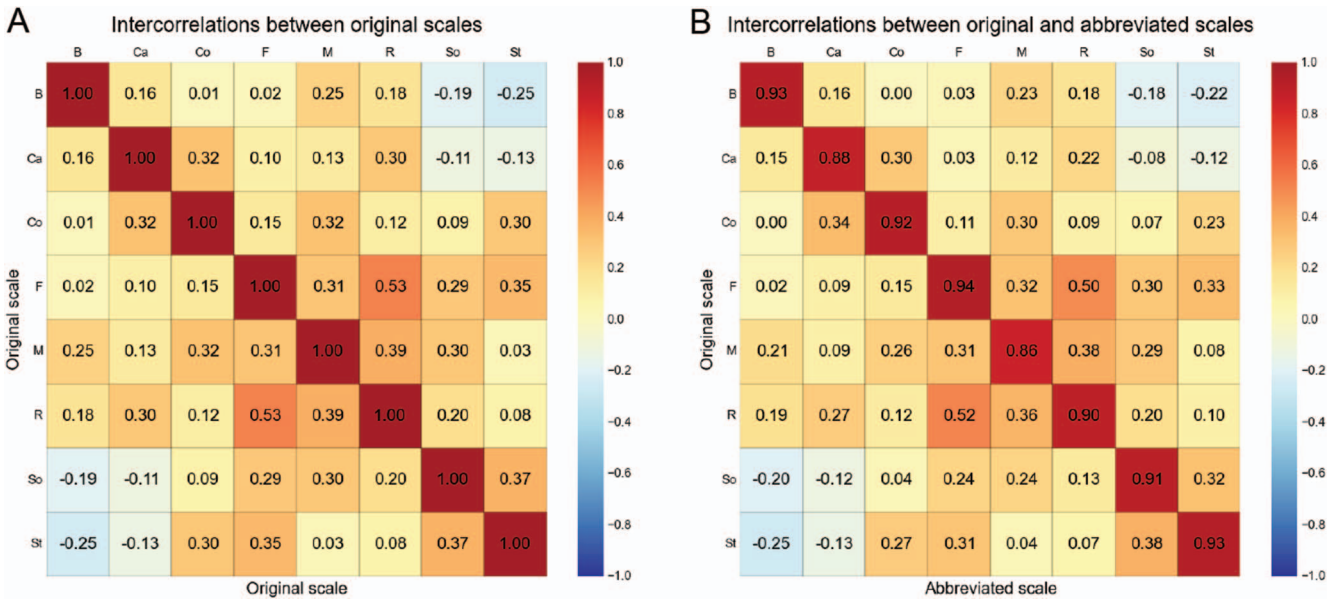

Figure 1. (A) Intercorrelations between the eight PPI-R scales for the original (i.e., unabbreviated) measure. (B) Correlations between the abbreviated PPI-R scales (*x*-axis) and original PPI-R scales (*y*-axis). Note the similarity between the two matrices and the high convergent correlations (diagonal values in [B]), B = Blame Externalization; Ca = Carefree Nonplanfulness; Co = Coldheartedness; *F* = Fearlessness; M = Machiavellian Egocentricity; R = Rebellious Nonconformity; So = Social Influence; St = Stress Immunity. See the online article for the color version of this figure.

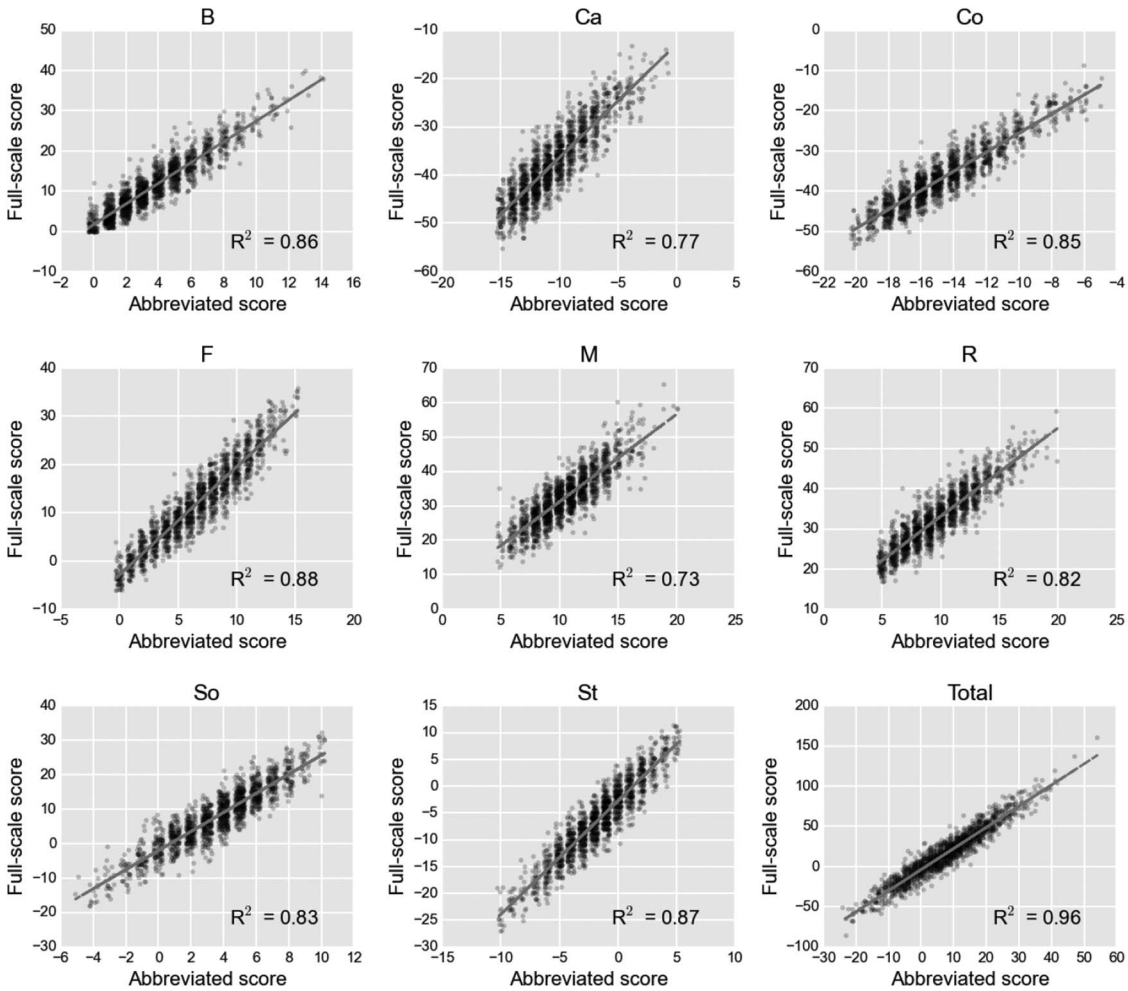

Figure 2. Scatterplots displaying convergent correlations between the original and abbreviated PPI-R measures for the eight PPI-R scales as well as the total PPI-R score (bottom right). Scores are randomly jittered slightly (+/−0.3 units drawn from a uniform distribution) on the x and y axes in order to prevent banding, B = Blame Externalization; Ca = Carefree Nonplanfulness; Co = Coldheartedness; *F* = Fearlessness; M = Machiavellian Egocentricity; R = Rebellious Nonconformity; So = Social Influence; St = Stress Immunity.

lation with each subscale. However, direct comparison of the GA-based measure with a "top-5" version indicated that the former measure performed substantially better. When each PPI-R subscale was scored using the top five items from the original version, the mean R^2 was .79 (range = .70–.86), a substantial reduction from the GA-based version (M = .83). Strikingly, the PPI-R-40 retained an average of only 3.1 of the top five individual items from each subscale, demonstrating its capacity to programmatically discard items that displayed high zero-order correlations with subscale scores but that were highly redundant with other items.

Study 2

Study 1 demonstrated that our automated approach was capable of producing an abbreviated version of the PPI-R that considerably shortened the length of the measure with high fidelity. In, Study 2, we tested the PPI-R-40's ability to generalize to a new sample, as well as its capacity to recapture PPI-R associations with external variables. We correlated the subscale scores and the sum score of the abbreviated and the original length versions in an additional independent sample (n = 206). As can be seen in Table 3, correlations again varied between r = .83 and .95 (M = .90) with a correlation of .89 for Self-Centered Impulsivity, of .94 for Fearless Dominance and of .93 for the sum score, again demonstrating that the PPI-R-40 maintained convergent validity in entirely new samples. In addition, the PPI-R-40 showed very similar correlations to the original PPI-R with the ZKPQ impulsivity subscale and the SPSRQ subscales. Using Fischer's z test for the significance of the difference between dependent correlations, the PPI-R/PPI-R-40 correlations did not significantly differ (see p scores for the z test in Table 3).

Study 3

A potential limitation of the PPI-R-40 was that it was developed using a German translation of the PPI-R. Although there is prom-

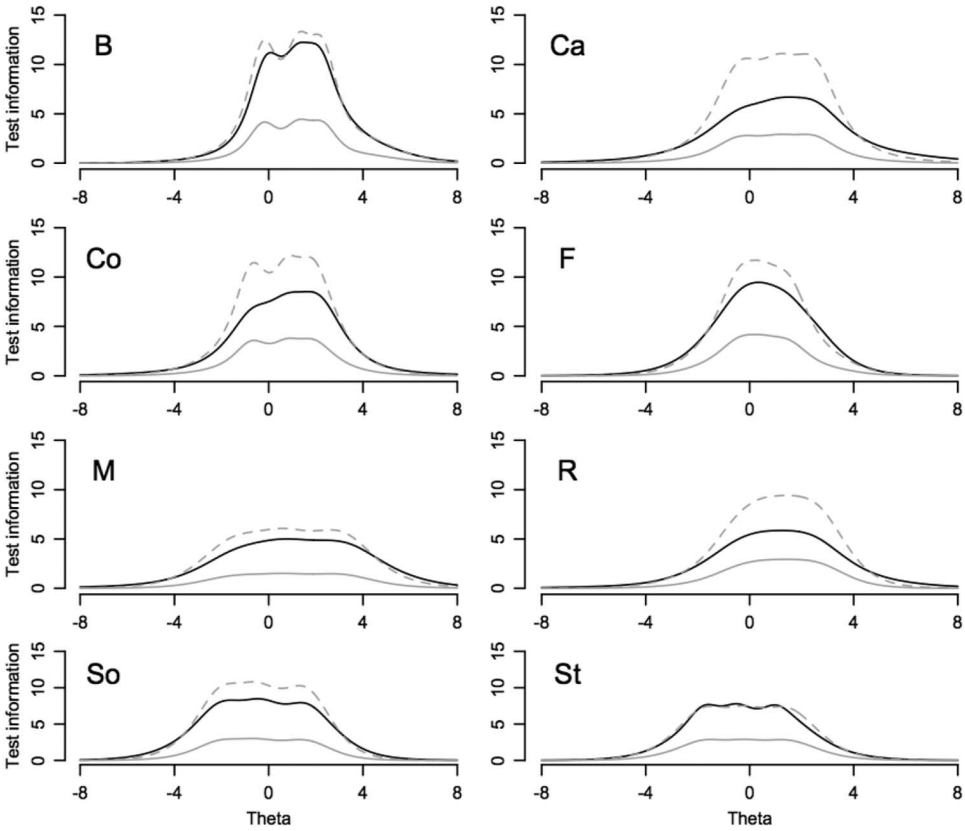

Figure 3. Total IRT test information for each scale of the original and abbreviated PPI-R measures. Black line: original scale; solid gray line: abbreviated scale; dashed gray line: abbreviated scale adjusted for reduced number of items.

ising construct validity for this this translation, which shows very similar properties to the original English version (Eisenbarth & Alpers, 2007), it was important to establish the generalizability of the PPI-R-40 to English samples as well. To accomplish this goal, in Study 3 we tested the PPI-R-40 on data acquired from an English-speaking sample using the English version of the PPI-R (n = 229). We used the PPI-R-40 scoring key generated in Study 1 to compute abbreviated scores for the English PPI-R. As in Studies 1 and 2, the resulting scores were highly correlated with the original (full-scale) scores (subscales: M = .91, range = .85 to .96; sum score: .95; Fearless Dominance: .96, Self-Centered Impulsivity: .90). The internal consistency (Cronbach's alpha) of the subscales in the sample ranged between .57 and .83 (M = .72), with an alpha of .81 for the sum score, .84 for the factor score of

Table 3
Correlations of PPI-R Subscales of the Abbreviated Version With the Subscales of the Original Version in Study 2 and With ZKPQ and SPSRQ and Significance of Fischer's Z-Test (p) Comparing the Correlation Coefficients Based on Sample 4 (N = 206)

	PPI-R original version	ZKPQ abb./original (p)	SPSRQ abb./original (p) P	SPSRQ abb./original (p) R
Blame Externalization	.92	.11/.06 (.61)	.37/.36 (.91)	.18/.15 (.76)
Carefree Nonplanfulness	.86	.23/.28 (.59)	−.04/.04 (.42)	−.04/.01 (.61)
Coldheartedness	.88	−.22/−.13 (.35)	−.11/−.23 (.21)	.02/.04 (.84)
Fearlessness	.93	.53/.59 (.38)	−.25/−.32 (.44)	.29/.34 (.58)
Machiavellian Egocentricity	.83	.21/.16 (.60)	−.02/.01 (.76)	.52/.54 (.78)
Rebellious Nonconformity	.90	.67/.69 (.71)	−.16/−.15 (.92)	.30/.31 (.91)
Social Influence	.90	.24/.34 (.27)	−.52/−.57 (.47)	.29/.35 (.50)
Stress Immunity	.95	.15/.13 (.84)	−.52/−.55 (.67)	.01/.03 (.84)
Self-centered Impulsivity	.89	.30/.25 (.59)	.18/.20 (.83)	.34/.35 (.91)
Fearless Dominance	.94	.46/.51 (.51)	−.60/−.67 (.24)	.29/.36 (.43)
Sum Score	.93	.59/.60 (.88)	−.37/−.40 (.72)	.44/.49 (.52)

Note. P = Sensitivity to punishment; R = Sensitivity to reward.

Fearless Dominance and .75 for the factor score of Self-Centered Impulsivity.

Discussion

The goal of this series of studies was to generate and provide preliminary validation data on an abbreviated version of the Psychopathic Personality Inventory–Revised using the automated approach recently introduced in Yarkoni (2010). The abbreviated measure, the PPI-R-40, showed highly convergent correlations in a series of unbiased analyses spanning several independent samples. In addition, the PPI-R-40 considerably outperformed an alternative abbreviated measure generated using the more conventional approach of retaining the top *N* items for each scale (Goldberg et al., 2006). The brevity of our measure relative to the original PPI-R (40 vs. 154 items) provides researchers with considerable time savings when assessing psychopathic traits, with relatively little loss of fidelity to the original measure.

Our research provides further evidence for the flexibility and utility of the GA method to shorten measurements. The 40-item PPI-R-40 strikes what is, in our view, an ideal balance between brevity and fidelity; however, as illustrated in Table 1, the length of a measure, and the amount of redundancy it allows (i.e., the degree to which individual items can be used to score multiple scales) can be easily adjusted depending on the specific needs of the researcher. Thus, the development of an abbreviated measure using the present approach is driven primarily by pragmatic considerations rather than by conventional psychometric criteria such as internal consistency or factor structure (see Smith, McCarthy, & Anderson, 2000). Indeed, low internal consistency in the sense of low homogeneity (when coupled with high concurrent and predictive validity) is the defining property of a successfully abbreviated measure, as high internal consistency values would imply residual redundancy between items (for a detailed discussion see Yarkoni, 2010). As we demonstrate in Studies 1–3, the PPI-R-40 shows high convergence with the PPI-R in terms of its ability to recapture rank-order score distributions, interscale correlations, and correlations with external measures (including correlations with the SPSR and Zuckerman Scales that are as strong as those obtained for the full PPI-R). Crucially, the PPI-R-40 continues to perform well even when tested in a translated (English) version using data acquired in a very different setting (Study 3).

Despite the clear validity and practical utility of the PPI-R-40, several limitations are worth noting. First, the current GA method is based on classical measurement theory rather than newer measurement models such as item response theory. The reliance on a classical measurement model may limit the PPI-R-40's ability to discriminate between individuals at the extremes of the distribution, which could be especially important in clinical or forensic samples. However, this is not a principled limitation, as our open-source tools (http://github.com/tyarkoni/scythe) could be readily adapted to yield a cost function based on IRT in future research (though we note that such an approach would present computational challenges, as it would require fitting hundreds of thousands of polytomous IRT models). In the present study, the information function for the PPI-R-40 closely resembled that of the full-length PPI-R (see Figure 3), suggesting that the potential benefits of IRT in this context are likely to be modest.

Second, strictly speaking, the abbreviated measures generated using our approach—including the PPI-R-40—will rarely if ever be optimal. Rerunning the genetic algorithm will produce a somewhat different abbreviation each time—the vast majority of which will nonetheless perform comparably according to the evaluation metric. However, this is largely a consequence of the high dimensionality of the space, which precludes an exhaustive search of all possible abbreviations—for example, there are over 2^{37} ways to select 40 items from a pool of 156. Moreover, in most cases, contextual factors (many of which are difficult to objectively quantify) are likely to play a much greater role in determining what constitutes an "ideal" abbreviated measure than will the actual results of any search algorithm. For example, no matter how well a measure such as the PPI-R-40 performs on a given set of objective metrics, it will be clearly suboptimal for an investigator who, say, only has time to administer 20 items, or who needs to maximize fidelity on just one or two PPI-R scales rather than all eight, and so on. The twofold benefit of the approach we adopt here is that (a) it allows us to produce a "good enough" measure such as the PPI-R-40 that is likely to perform very well across a broad range of common applications, and (b) the same programmatic approach can be used to easily generate alternative forms of the PPI-R (or other measures) in cases where investigators have more idiosyncratic needs. Lest one worry that this flexibility opens the door to a potential proliferation of different versions of the same measure, we note that, in practice, the similarity between any two abbreviations generated using our approach is likely to be far greater than the similarity between putative alternative forms of most other measures, or between different measures of the same putative construct. Put differently, any researcher who is comfortable treating, say, the 60-item NEO-FFI and the 240-item NEO-PI-R as if they measure the same construct of Extraversion should have no compunction about treating different version of the PPI-R as functionally equivalent as well.

Third, scores produced by the PPI-R-40 are not commensurable with existing norms for the full PPI-R. Thus, the PPI-R-40 cannot be used to screen population samples based on previously established cut-offs. However, such applications are extremely rare; in practice, the vast majority of studies use the PPI-R in a correlational fashion, and Studies 1–3 demonstrate that the PPI-R-40 is an excellent substitute in this latter regard. Fourth and finally, like all construct validation endeavors (Cronbach & Meehl, 1955), our set of studies can be viewed as a work in progress. Further work using additional measures of constructs relevant to psychopathy (e.g., callousness, lack of guilt) as well as different modes of assessment, including external criteria (e.g., interview, laboratory measures) relevant to psychopathy will be needed to further establish the comparability of the PPI-R-40 with the PPI-R. Nevertheless, the findings of this initial investigation are extremely encouraging. In sum, our results introduce a new instrument for high-efficiency measurement of psychopathic traits to the literature, while providing additional evidence of the utility of automated abbreviation of personality measures.

References

Alpers, G. W., & Eisenbarth, H. (2008). *Psychopathy Personality Inventory Revised–Deutschsprachige version. Testhandbuch (Test guide).* Göttingen, Germany: Hogrefe.

Benning, S. D., Patrick, C. J., Hicks, B. M., Blonigen, D. M., & Krueger, R. F. (2003). Factor structure of the psychopathic personality inventory: Validity and implications for clinical assessment. *Psychological Assessment, 15,* 340–350. http://dx.doi.org/10.1037/1040-3590.15.3.340

Cronbach, L. J. (1954). Report on a psychometric mission to clinicia. *Psychometrika, 19,* 263–270. http://dx.doi.org/10.1007/BF02289226

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Eisenbarth, H., & Alpers, G. W. (2007). Validierung der deutschen Übersetzung des Psychopathy Personality Inventory (PPI). *Zeitschrift für Klinische Psychologie und Psychotherapie, 36,* 216–224. http://dx.doi.org/10.1026/1616-3443.36.3.216

Fulton, J. J., Marcus, D. K., & Payne, K. T. (2010). Psychopathic personality traits and risky sexual behavior in college students. *Personality and Individual Differences, 49,* 29–33. http://dx.doi.org/10.1016/j.paid.2010.02.035.

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. http://dx.doi.org/10.1016/j.jrp.2005.08.007

Hare, R. D. (2003). *Manual for the Hare Psychopathy Checklist–Revised* (2nd ed.). Toronto, ON: Multi-Health Systems.

Hewig, J., Hagemann, D., & Riemann, R. (2014). Translated version of Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ). Manuscript in preparation.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a non-institutionalized population. *Journal of Personality and Social Psychology, 68,* 151–158.

Lilienfeld, S. O., & Widows, M. R. (2005). *Psychopathy Personality Inventory Revised (PPI-R). Professional manual*. Lutz, FL: Psychological Assessment Resources.

Lykken, D. T. (1995). *The antisocial personalities*. Hillsdale, NJ: Erlbaum.

Neumann, C. S., Malterer, M. B., & Newman, J. P. (2008). Factor structure of the Psychopathic Personality Inventory (PPI): Findings from a large incarcerated sample. *Psychological Assessment, 20,* 169–174. http://dx.doi.org/10.1037/1040-3590.20.2.169

Ostendorf, F., & Angleitner, A. (1994). A comparison of different instruments proposed to measure the Big Five. *European Review of Applied Psychology, 44,* 45–53.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411–419.

Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy: Developmental origins of disinhibition, boldness, and meanness. *Development and Psychopathology, 21,* 913–938. http://dx.doi.org/10.1017/S0954579409000492

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111. http://dx.doi.org/10.1037/1040-3590.12.1.102

Tonnaer, F., Cima, M., Sijtsma, K., Uzieblo, K., & Lilienfeld, S. (2013). Screening for psychopathy: Validation of the Psychopathic Personality Inventory–Short Form with Reference Scores. *Journal of Psychopathology and Behavioral Assessment, 35,* 153–161.

Torrubia, R., Avila, C., Molto, J., & Caseras, X. (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. *Personality and Individual Differences, 31,* 837–862. http://dx.doi.org/10.1016/S0191-8869(00)00183-5

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality, 44,* 180–198. http://dx.doi.org/10.1016/j.jrp.2010.01.002

Zuckerman, M. (2002). Zuckerman–Kuhlman Personality Questionnaire (ZKPQ): An alternative five-factorial model. In B. De Raad & M. Perugini (Eds.), *Big Five assessment* (pp. 377–396). Seattle, WA: Hogrefe and Huber Publishers.

Received February 9, 2014
Revision received August 20, 2014
Accepted September 3, 2014 ■