Condon, D M and Revelle, W 2015 Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items. *Journal of Open Psychology Data,* 3: e6, DOI: http://dx.doi.org/10.5334/jopd.al

DATA PAPER

Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items

David M. Condon[1] and William Revelle[2]

[1] Department of Medical Social Sciences, Northwestern University, Chicago, Illinois
david-condon@northwestern.edu

[2] Department of Psychology, Northwestern University, Evanston, Illinois
revelle@northwestern.edu

These data were collected to evaluate the structure of personality constructs in the temperament domain. In the context of modern personality theory, these constructs are typically construed in terms of the Big Five (Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion) though several additional constructs were included here. Approximately 24,000 individuals were administered random subsets of 696 items from 92 public-domain personality scales using the Synthetic Aperture Personality Assessment method between December 8, 2013 and July 26, 2014. The data are available in rdata and csv formats and are accompanied by documentation stored as a text file. Re-use potential includes many types of structural and correlational analyses of personality.

Keywords: Personality; temperament; personality structure; Big Five
Funding statement: This research was supported in part by grant SMA-1419324 from the National Science Foundation to William Revelle and David Condon.

1 Overview

Collection Date(s)

Data were collected between December 8, 2013 and July 26, 2014.

Background

The SAPA Project is a collaborative online data collection tool for assessing psychological constructs across multiple domains of personality. These domains – temperament, cognitive abilities, and interests – have been chosen based on historical and current prominence in the field of individual differences research. The primary goal of the SAPA Project is to determine the combined and independent structures of each of these domains based on the collection of large, cross-sectional, online samples. Secondary goals include (1) the identification of additional domains (e.g., motivation, character) which may also provide insight into the ways that individuals differ; and (2) an improved understanding of the demographic and psychographic correlates of individual differences in personality.

The data described here were collected in order to evaluate the structure of personality constructs in the temperament domain. In the context of modern personality theory, these constructs are typically construed in terms of the Big Five (Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion). While several large scale studies of personality have been conducted using trait descriptive adjectives and nouns (see [6] and [11]), relatively few attempts have been made to evaluate large sets of phrased items even though phrased item types are more typically used in personality assessments.

Items from 92 public-domain personality scales were included in this data set; most of these were chosen explicitly because they are among the more widely-used personality measures. No a priori hypotheses were made regarding the underlying structure of these items, nor should it be expected that these items represent an unbiased or representative snapshot of human personality; the structure of these items will, to some extent, reflect the shared characteristics of the scales from which they were taken.

2 Methods

Sample

Participants (N = 23,681) completed the online survey in exchange for feedback about various aspects of their personality. No active advertisements or marketing efforts were used to attract participants for this data collection; web traffic statistics (collected through Google Analytics) suggest that participants who did not come to the website directly were directed to it through links from various other websites about personality, personality research, general psychology topics, and psychometrics. Many of these websites were academic/educational in nature.

Many demographic and psychographic variables are included in the data. These include: gender (64% of the participants were female); age (see **Figure 1**); marital status (see **Table 1**); body mass index (see **Figure 2**); country (172 countries were represented in total; 69.5% of participants were from the United States, and 12 countries had more than 100 participants); state/region (for 32 countries); zip code (postal codes for U.S. participants only); race/ethnicity (see **Table 2**); educational attainment level (see **Table 3**); employment status (see **Table 4**); parental educational attainment level (for 1 or 2 parents); and parental field of employment (for 1 or 2 parents). Participants were not required to provide any of these data except age and gender.

Materials

Items from eight sets of self-report personality scales were administered. Seven of these are based on items from the International Personality Item Pool [7; 9]: the 100 IPIP items corresponding to the Big Five factor markers [7], the 100 items of the Big Five Aspect Scales [3], the 240 items of the IPIP-HEXACO inventory [1], the 48 items of the Questionnaire Big Six scales [18], the 300 items of the IPIP-NEO [7], the 127 items of the IPIP-Multidimensional Personality Questionnaire ["MPQ" 8; 17], and the 40 items of the Plasticity/Stability scales [2]. The eighth set of

Marital Status	Participants
Never Married	14,197
Married	3,736
Domestic Partnership	1,103
Widowed & Remarried	769
Divorced & Single	300
Divorced & Remarried	123
Widowed & Single	14

Table 1: Participants by Marital Status.

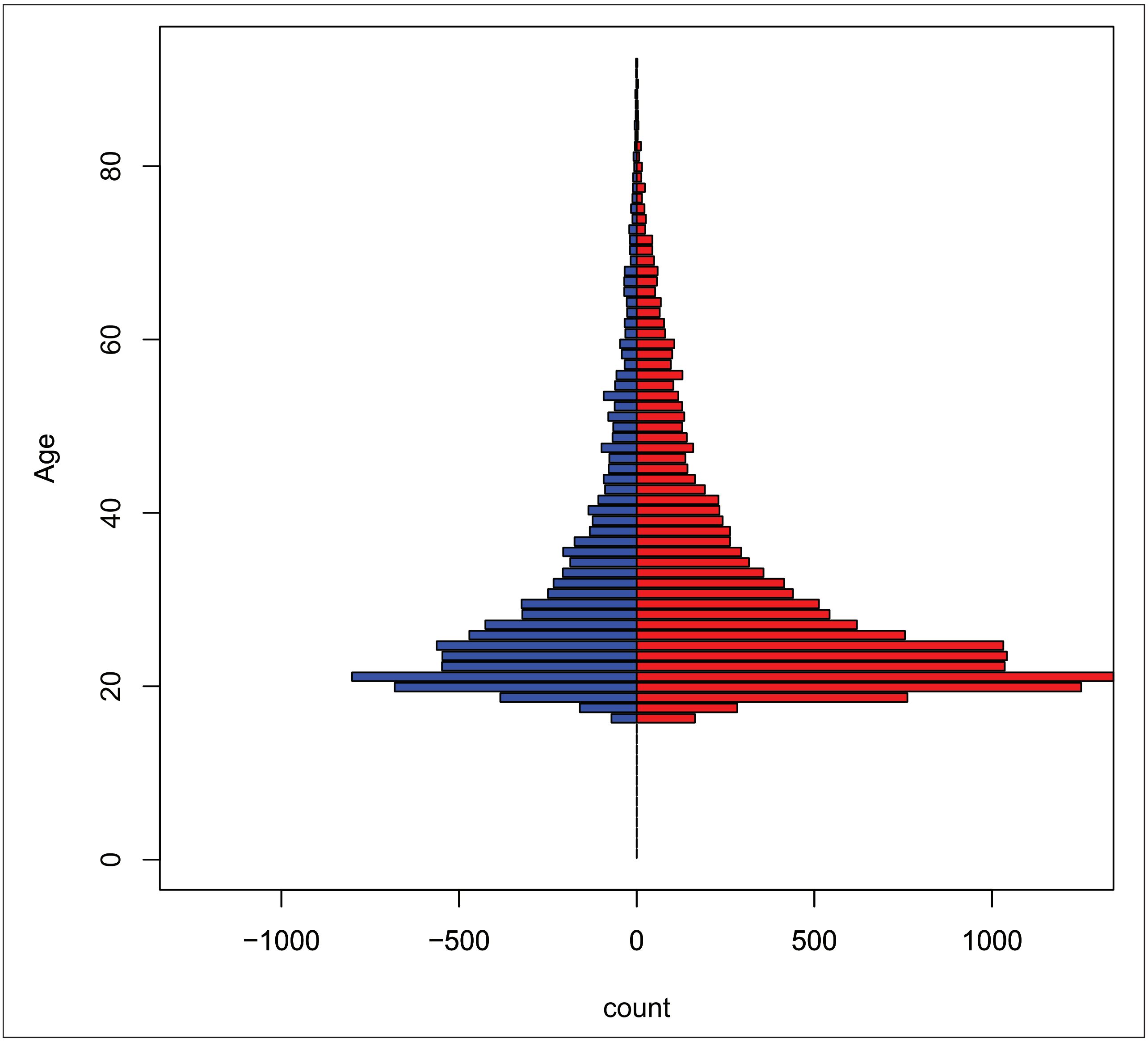

Figure 1: Participants by Marital Status.

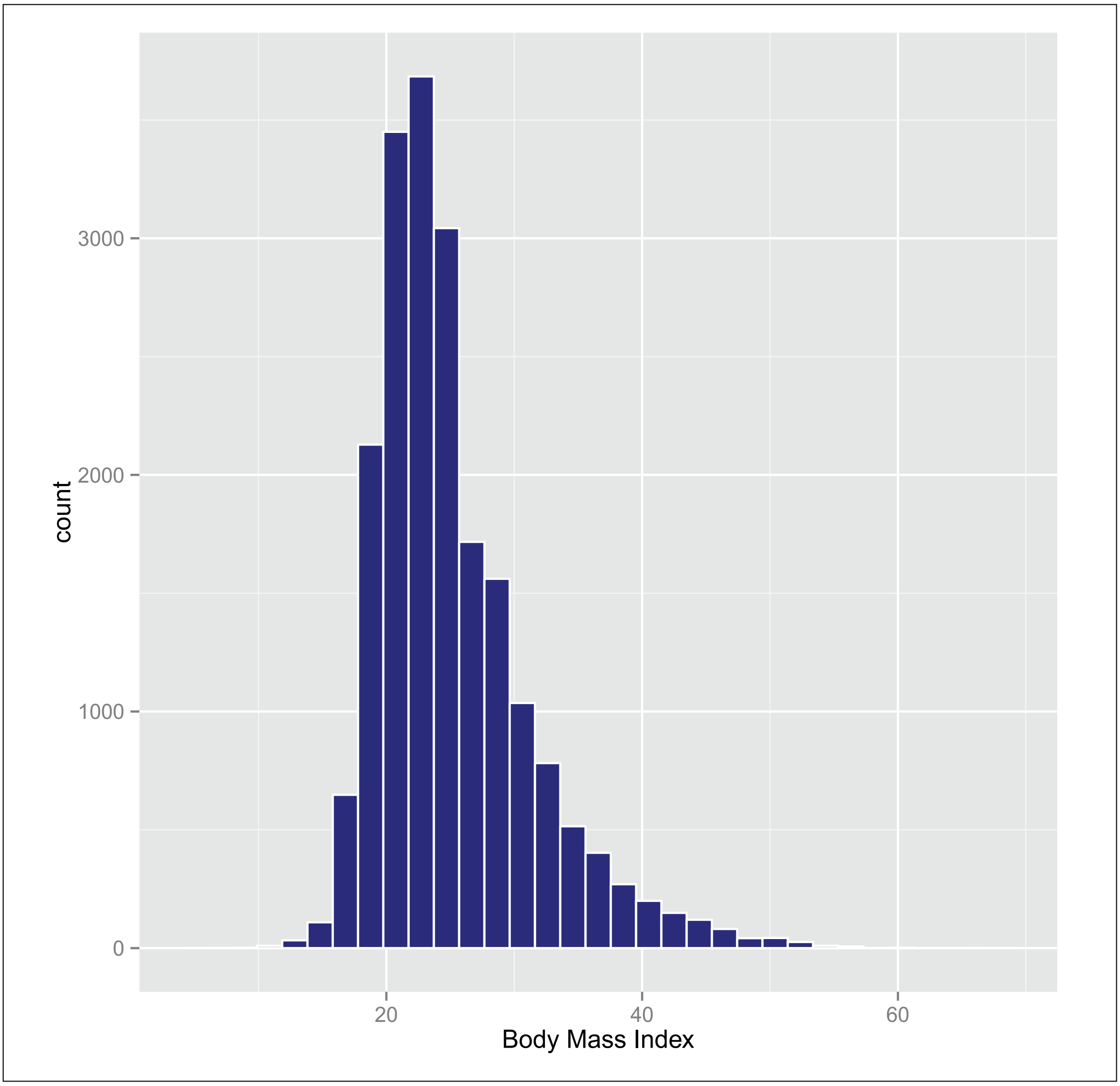

Figure 2: Participants by Body Mass Index.

scales included 79 items from the Eysenck Personality Questionnaire – Revised [5]. Note that the format of these items was modified to match that of the IPIP items and that the 21 "lie" scale items were intentionally omitted. Administration of these scales also implies the administration of several other measures which are abbreviations of these scales, including the 24 and 36 item Questionnaire Big Six scales [18], the 50 item IPIP scales corresponding to the Big Five factor markers [8], and the 20 item "mini-IPIP" scales [4].

The 1,034 items from these measures contain 338 duplicates, resulting in a total set of 696 unique items. Of these, 473 items are in only one set of scales, 126 items are included in two sets of scales, 54 items are in three, 22 items in four, 17 items in five, and 4 items are in six of the seven sets of IPIP-based scales ("Have little to say", "Worry about things", "Like order", and "Have a rich vocabulary"). All of the items were administered with the same six response options ("Very Inaccurate", "Moderately Inaccurate", "Slightly Inaccurate", "Slightly Accurate", "Moderately Accurate", "Very Accurate").

Procedures

The items were administered using the Synthetic Aperture Personality Assessment ("SAPA") technique [16], a variant of matrix sampling procedures discussed by Lord [12]. This method produces data which contain "massive missingness" by design [15]. This missingness qualifies for classification as missing completely at random ["MCAR", 10] and it is further described as massively missing because the mean level of missingness by participant was approximately 86%. The personality items were presented to participants in random order, and participants responded to as many items as they wished. The mean number of personality items to which participants responded was 86.1 (*sd* = 58.7;

Ethnicity	Participants
White	8,291
African American	1,329
Two Or More Ethnicities	809
Mexican	687
Other Hispanic	434
Chinese	241
Other Asian	191
Indian	128
Puerto Rican	128
Filipino	119
Native American	100
Korean	73
Cuban	35
Japanese	23
Other Pacific Islander	22
Native Hawaiian	9
Alaskan Native	6
Other	185

Table 2: Race/Ethnicity Among U.S. Participants.

Educational Level	Participants
Less than 12 years	2,725
High school graduate	1,913
Currently in college/university	8,457
Some college/university, but did not graduate	1,249
College/university degree	3,114
Currently in graduate or professional school	1,108
Graduate or professional school degree	2,007

Table 3: Participants by Educational Attainment Level.

median = 71). The number of items administered to each participant was procedurally independent of participant response characteristics. Participants were encouraged to complete approximately 100 items but were able to complete up to 330. The number of administrations for each item varied considerably (median = 2554; *m* = 2931; *sd* = 781) as did the number of pairwise administrations between any two items in the set (median = 519; *m* = 528; *sd* = 117).

Quality Control

The available data are presented largely as they were collected with only two exceptions:

1. Partial removal of data collected from participants who completed the survey more than once in a single browser session. This was done by assigning participants a random user ID that was persistant as long as their current browser session remained active. In those cases where more than 1 response set was entered in a single browser session, only the first response set was kept.
2. Removal of participants with self-reported ages younger than 14 and older than 90. The survey is not intended for participants younger than 14. Self-reported ages over 90 were removed on the grounds that they were deemed to be unlikely.

Employment Status	Participants
Employed	9,414
Currently a student	8,073
Not employed	1,128
Not employed, seeking work	892
Homemaker	341
Retired	236

Table 4: Participants by Employment Status.

Ethical issues

No personally identifying information were collected from participants in these data.

3 Dataset description

Object name

For use with R statistical software systems [13], the data file is named: 'sapaTempData696items08dec2013thru26jul2014.rdata'

The data file is named to indicate the domain (temperament), the number of items included (696), and the time period over which the data were collected (08dec2013 through 26jul2014). The file can be found at: http://dx.doi.org/10.7910/DVN/SD7SVE

The data file includes four objects. The most pertinent of these is the raw data object ('sapaTem-pData696items08dec2013thru26jul2014'). The remaining three objects are helper files for data analysis: 'ItemInfo696' is a data dictionary which provides the text for each of the temperament items and a listing of the scales with which it is associated, 'ItemLists' is a list object that provides an index of all the temperament items associated with each measure, and 'superKey696' is a scoring matrix for the many personality scales which can be scored based on these data.

The data are also available in a csv format, available at the same location listed above. Each of the four objects described above has been saved as a separate csv file.

Data type

Self-report, cross-sectional survey data from 23,681 participants.

Format names and versions

The data are stored as an rdata file (approximately 4.9 MB). This file includes the four objects described above: the main data object, 'sapaTempData696items08dec2013thru26jul2014', as well as 'ItemInfo696', 'ItemLists', and 'superKey696'. It should be noted that several of the scales in these measures require reverse coding of some items; see the original documentation of each measure for more details.

In addition to the rdata file containing these four objects, there is also an associated text file that provides full information on the demographic codes ('demographic codes.txt').

Data Collectors

In addition to the authors, Jason French, Lorien Elleman and Zara Wright contributed by helping to maintain the website and increase its visibility.

Language

All aspects of the survey and website were written in English. Data collected about the website through Google Analytics suggests that some participants used browser-based translation software, but no specifics are available about the extent and effect of these translations.

License

The data have been deposited under the open license CC0 (Public Domain Dedication).

Embargo

The data are freely available for use with appropriate citation.

Repository location

The data were published on Dataverse and are located at http://dx.doi.org/10.7910/DVN/SD7SVE.

Publication date

The dataset was published on 13/04/2015.

4 Reuse potential

The data are well-suited for many types of structural and correlational analyses of personality. These might include evaluations based on one of the many measures independently, the ways in which these measures relate to one another, exploratory evaluations of their shared structure, and evaluations of structural relationships across constructs in various groups of participants (e.g., based on age, gender, country/region, educational attainment levels, etc.). The large number of both participants and items also make it possible to construct novel scales based on the empirical correlations between items and criterion variables (see the 'bestScales' function and related help pages in the *psych* package [14] for examples of these techniques). Additional, non-overlapping data sets from the SAPA Project are also available for use; contact the authors for more information.

Competing Interests

The authors declare that they have no competing interests.

Paper Author Contribution and Affiliations

The first author is David M. Condon, PhD, an Assistant Professor at Northwestern University's Feinberg School of Medicine in the Department of Medical Social Sciences. His contribution included a primary role in the technical development of the website through which these data were collected (sapa-project.org). This involved the adaptation of existing code (primarily generated by the second author for previous data collection projects) and the authorship of new code (for extending the functionality and aesthetic design of the website and for improving the data storage and data exportation methods). The first author also took the lead role in cleaning the data to prepare it for sharing, making the data available in the Dataverse, and preparing and submitting this manuscript.

The second author, William Revelle, PhD, is credited with first implementing the survey sampling techniques that made this data collection possible (Synthetic Aperture Personality Assessment, "SAPA") and for developing earlier incarnations of the survey described on his website (personality-project.org). He also owns the url for the website where these data were collected (sapa-project.org). The second author played a secondary role in the development and maintenance of the website used to collect these data.

Acknowledgements

The authors would like to acknowledge Joshua Wilt for his contributions on prior incarnations of the SAPA-Project.

References

1. **Ashton, M C, Lee, K** and **Goldberg, L R** 2007 "The IPIP–HEXACO scales: An alternative, public-domain measure of the personality constructs in the HEXACO model". *Personality and Individual Differences*, 42(8): 1515–1526. DOI: http://dx.doi.org/10.1016/j.paid.2006.10.027
2. **DeYoung, C G** 2010 "Toward a Theory of the Big Five". *Psychological Inquiry*, 21(1): 26–33. DOI: http://dx.doi.org/10.1080/10478401003648674
3. **DeYoung, C G, Quilty, L C** and **Peterson, J B** 2007 "Between facets and domains: 10 aspects of the Big Five". *Journal of Personality and Social Psychology*, 93(5): 880–896. DOI: http://dx.doi.org/10.1037/0022-3514.93.5.880. PMid: 17983306.
4. **Donnellan, M B, Oswald, F L, Baird, B M** and **Lucas, R E** 2006 "The Mini-IPIP Scales: Tiny-yet-effective measures of the Big Five Factors of Personality". *Psychological Assessment*, 18(2): 192–203. DOI: http://dx.doi.org/10.1037/1040-3590.18.2.192. PMid: 16768595.
5. **Eysenck, S H, Eysenck, S** and **Barrett, P** 1985 "A revised version of the psychoticism scale". *Personality and Individual Differences*, 6(1): 21–29. DOI: http://dx.doi.org/10.1016/0191-8869(85)90026-1
6. **Goldberg, L R** 1992 "The development of markers for the Big-Five factor structure". *Psychological Assessment*, 4(1): 26–42. DOI: http://dx.doi.org/10.1037/1040-3590.4.1.26
7. **Goldberg, L R** 1999 "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor Models". In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.) *Personality and Individual Differences*, Tilburg University Press, The Netherlands, pp. 7–28.

8. **Goldberg, L R** 2014 "*International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*". Retrieved from http://ipip.ori.org/ [January 18, 2014].
9. **Goldberg, L R** and **Johnson, J A** 2006 "The international personality item pool and the future of public-domain personality measures". *Journal of Research in Personality*, 40(1): 84–96. DOI: http://dx.doi.org/10.1016/j.jrp.2005.08.007
10. **Graham, J W** 2009 "Missing Data Analysis: Making It Work in the Real World". *Annual Review of Psychology*, 60(1): 549–576. DOI: http://dx.doi.org/10.1146/annurev.psych.58.110405.085530. PMid: 18652544.
11. **John, O P** and **Srivastava, S** 1999 "The Big Five trait taxonomy: History, measurement, and theoretical perspectives". In L. Pervin & O. P. John (Eds.) *Handbook of personality: Theory and research*, Guilford Press, New York, pp. 102–138.
12. **Lord, F M** 1955 "Sampling fluctuations resulting from the sampling of test items". *Psychometrika*, 20(1): 1–22. DOI: http://dx.doi.org/10.1007/BF02288956
13. **R Core Team** 2015 "*R: A Language and Environment for Statistical Computing*". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
14. **Revelle, W** 2015 "psych: *Procedures for psychological, psychometric, and personality research*". Northwestern University, Evanston, Illinois, R package version 1.5.3.19.
15. **Revelle, W** and **Brown, A** 2013 "Standard errors for SAPA correlations". In *Society for Multivariate Experimental Psychology*, St. Petersburg, FL, pp. 1–10.
16. **Revelle, W, Wilt, J** and **Rosenthal, A** 2010 "Individual Differences in Cognition: New Methods for Examining the Personality-Cognition Link". In A. Gruszka, G. Matthews & B. Szymura (Eds.) *Handbook of Individual Differences in Cognition*, Springer New York, New York, NY, pp. 27–49. DOI: http://dx.doi.org/10.1007/978-1-4419-1210-7_2
17. **Tellegen, A** and **Waller, N G** 2008 "Exploring Personality Through Test Construction: Development of the Multidimensional Personality Questionnaire". In *The SAGE Handbook of Personality Theory and Assessment: Volume 2 — Personality Measurement and Testing*, SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP, United Kingdom, pp. 261–292. DOI: http://dx.doi.org/10.4135/9781849200479.n13
18. **Thalmayer, A G, Saucier, G** and **Eigenhuis, A** 2011 "Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires". *Psychological Assessment*, 23(4): 995–1009. DOI: http://dx.doi.org/10.1037/a0024165. PMid: 21859221.

Peer review comments: https://s3-eu-west-1.amazonaws.com/ubiquity-partner-network/up/journal/jopd/al/jopd-al.pdf

How to cite this article: Condon, D M and Revelle, W 2015 Selected Personality Data from the SAPA-Project: On the Structure of Phrased Self-Report Items. *Journal of Open Psychology Data* 3: e6, DOI: http://dx.doi.org/10.5334/jopd.al

Published: 18 August 2015

]u[*Journal of Open Psychology Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS