Condon, D M et al 2017 A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data,* 5: 3, DOI: https://doi.org/10.5334/jopd.32

DATA PAPER

A SAPA Project Update: On the Structure of phrased Self-Report Personality Items

David M. Condon[1], Ellen Roney[1] and William Revelle[2]

[1] Northwestern University, Department of Medical Social Sciences, US

[2] Northwestern University, Department of Psychology, US

Corresponding author: David M. Condon, PhD, Assistant Professor (david-condon@northwestern.edu)

Two large samples were collected to evaluate the structure of traits in the temperament domain. In both samples, participants were administered random subsets of public-domain personality items from a larger pool of approximately 700 items. These data broadly cover the most widely used, public-domain measures of personality (though this breadth is not likely free of theoretical bias). When combined with a third, previously-shared dataset that used the same methodological design [4], the sample includes more than 125,000 participants from more than 220 countries and regions. Re-use potential includes many types of structural, correlational, and network analyses of personality and a wide range of demographic and psychographic constructs. The data are available in both rdata and csv formats.

Keywords: Personality, personality structure, temperament, Big Five

(1) Overview

Collection Date(s)

Data were collected between July 26, 2014 and February 7, 2017 in two waves. The first was collected between July 26, 2014 and December 22, 2015 (denoted below as the "Replication Sample") and the second was collected between December 22, 2015 and February 7, 2017 (the "Confirmatory Sample"). These samples are related to a third dataset that was previously uploaded to Dataverse on July 6, 2015 and described in [4]. Those data were collected between December 8, 2013 and July 26, 2014. Collectively, these three samples represent approximately 1,150 days of uninterrupted cross-sectional survey data collection using the Synthetic Aperture Personality Assessment ("SAPA") technique.

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, values) 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 [9] and [15]), 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 approximately 400 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. They were not chosen based on any *a priori* hypotheses regarding the underlying structure of self-report 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.

It should be noted that the research design used to collect the two samples described herein (the Replication and Confirmatory Samples) was first described in the PhD dissertation of the first author [2], though these samples both have considerably more participants than the original sample (the Exploratory Sample). All three data sets were used to develop the SAPA Personality Inventory [3] and it is likely that other researchers who are interested in re-using these data would want to combine all three samples.

(2) Methods

Sample

Participants in both datasets (Replication Sample N = 54,855, Confirmatory Sample N = 48,350) completed an 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.

The data contain many demographic and psychographic variables. These include: gender (62% and 63% female in the Replication and Confirmatory Samples, respectively); age (see **Figures 1.1** and **1.2**); marital status (see **Figure 2**);

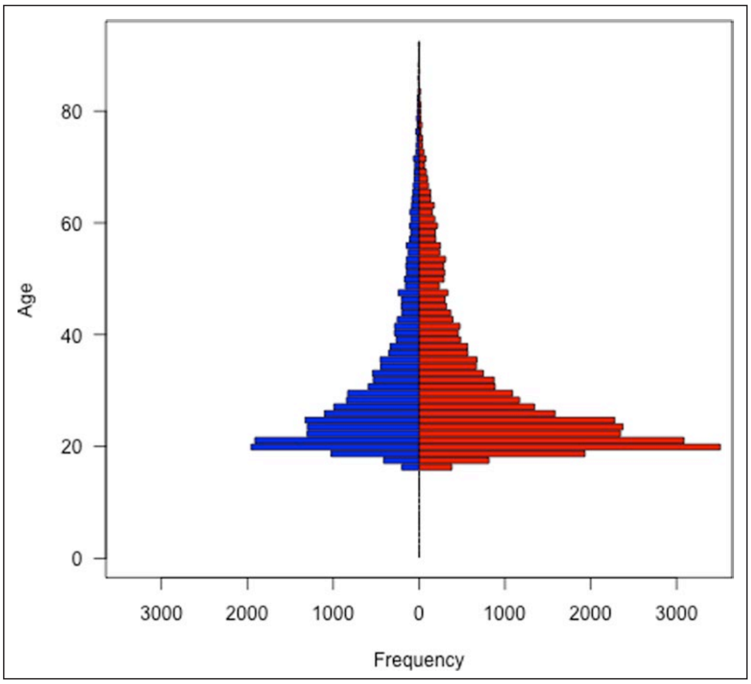

Figure 1.1: Participants by age and gender in the Replication Sample (males in blue, females in red).

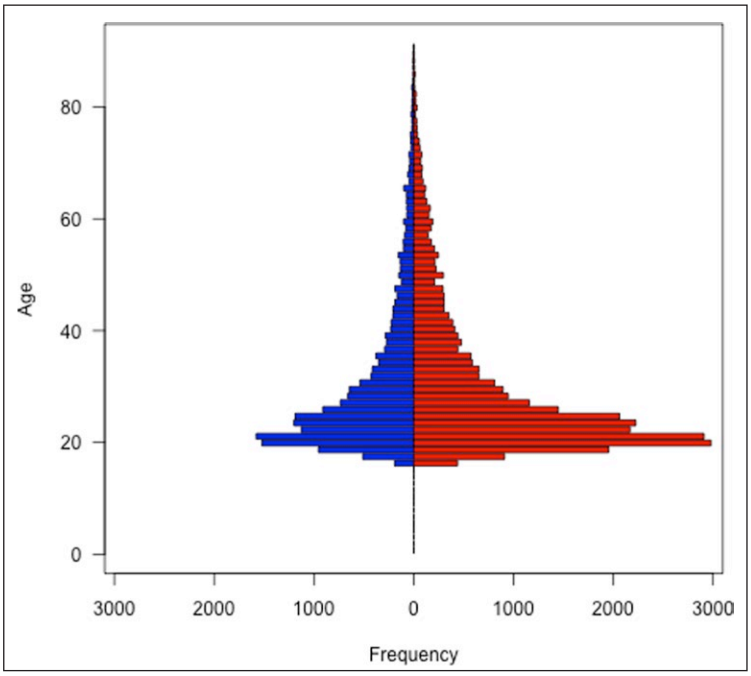

Figure 1.2: Participants by age and gender in the Confirmatory Sample (males in blue, females in red).

country (219 countries [14] were represented in the Replication Sample and 220 in the Confirmatory Sample; the number of countries with more than 50 participants was 50 and 48 in the Replication and Confirmatory samples, respectively; 68.7% and 66.5% of participants were from the United States); state/region (for 32 of the countries); ZIP code (postal codes for U.S. participants only); race/ethnicity (see **Figure 3**); educational attainment level (see **Figure 4**); employment status (see **Figure 5**); and parental field of employment (for 1 or 2 parents). Participants were not required to provide any of these data except age and gender.

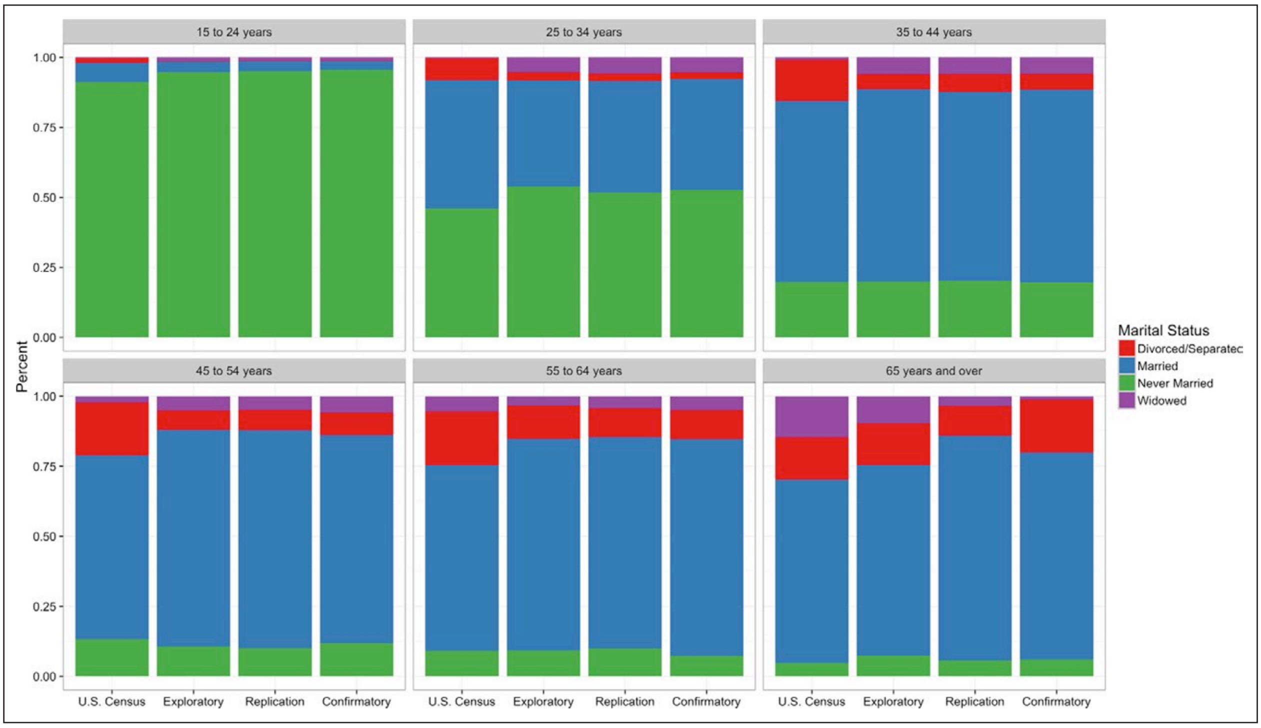

Figure 2: Distribution of marital status across ages in the U.S. Census and the Exploratory, Replication, and Confirmatory samples.

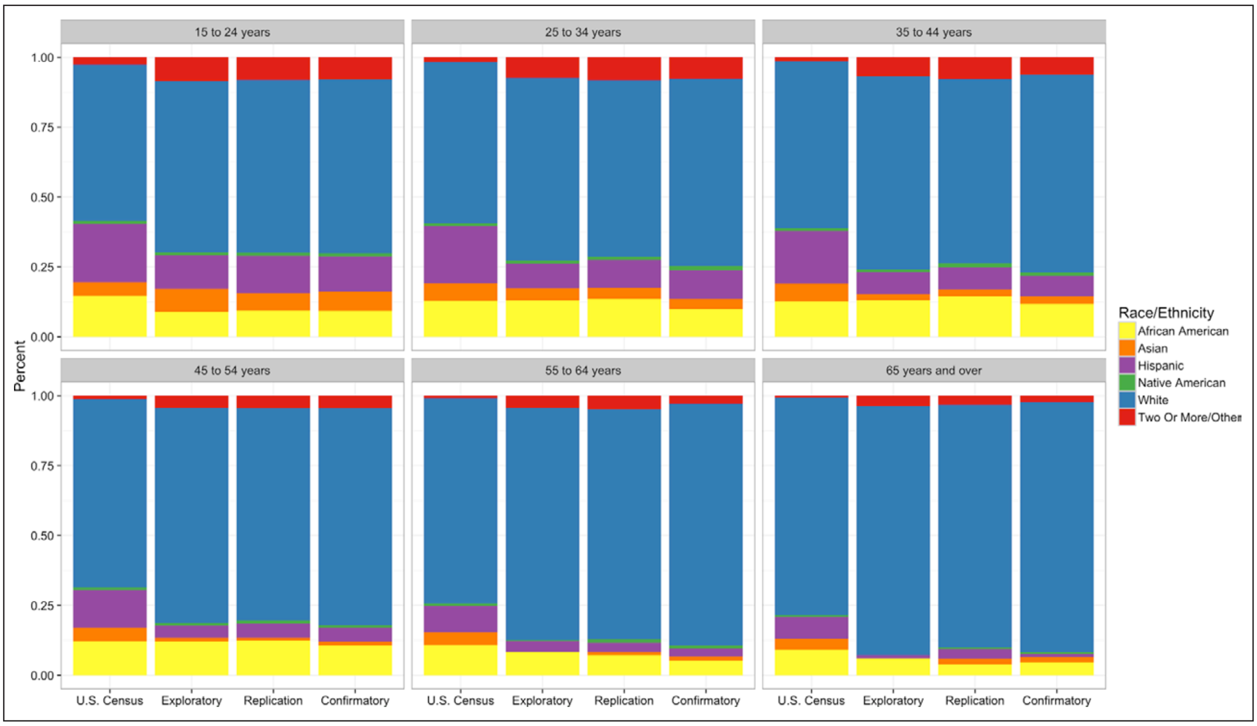

Figure 3: Distribution of race/ethnicity across ages in the U.S. Census and the Exploratory, Replication, and Confirmatory samples.

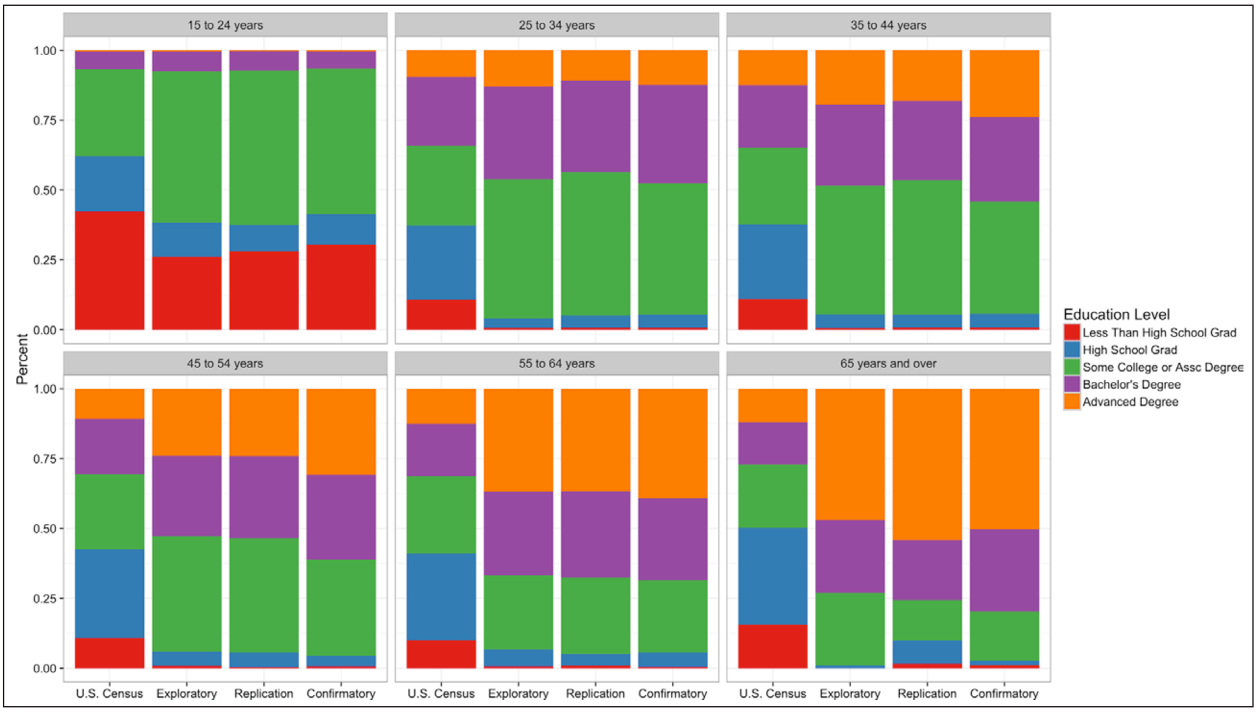

Figure 4: Distribution of educational attainment level across ages in the U.S. Census and the Exploratory, Replication, and Confirmatory samples.

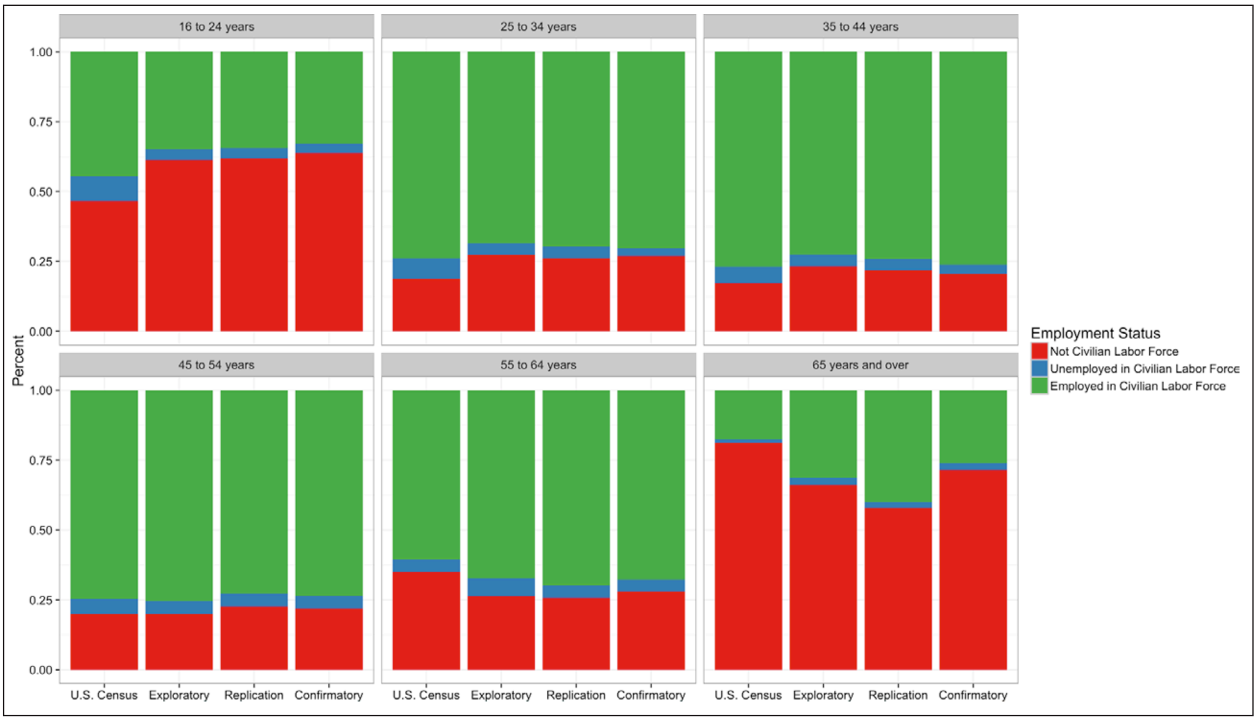

Figure 5: Distribution of employment status across ages in the U.S. Census and the Exploratory, Replication, and Confirmatory samples. Civilian labor force excludes anyone who is retired, a student, a homemaker, in jail, in an institution, or not seeking work.

Materials

All of the personality items administered in both samples are public-domain items and the majority of these are items from the International Personality Item Pool [10, 11, 12]. The items were primarily chosen based on their inclusion in at least one of eight sets of self-report personality scales. Seven of these eight sets of scales are available in the International Personality Item Pool, including: (1) the 100 IPIP items corresponding to the Big Five factor markers [9, 10]; (2) the 100 items of the Big Five Aspect Scales [6]; (3) the 240 items of the IPIP-HEXACO inventory [1]; (4) the 48 items of the Questionnaire Big Six scales [22]; (5) the

Table 1: Overview of the Size and Density of the Samples.

	Sample		
	Exploratory	**Replication**	**Confirmatory**
Total number of respondents	23,679	54,855	48,350
Mean number of responses per participant	86.1	84.9	84.7
Median number of responses per participant	71.0	71.0	71.0
Standard deviation of number of responses per participant	58.7	57.2	57.4
Mean percent of responses to all items (696)	13.9%	13.7%	13.6%
Median percent of responses to all items (696)	12.0%	12.0%	12.0%
Mean number of responses per item	2,931	6,693	5,881
Median number of responses per item	2,554	5,845	5,151
Standard deviation of number of responses per item	781	1,825.5	1,608.6
Mean pairwise responses for all possible pairings	528	1,180	1,037
Median pairwise responses for all possible pairings	519	1,155	1,014
Standard deviation of pairwise responses for all possible pairings	117	267	235

300 items of the IPIP-NEO [10]; (6) the 127 items of the IPIP-Multidimensional Personality Questionnaire [10, 21]; and (7) the 40 items of the Plasticity/Stability scales [5]. The eighth set of scales included 79 items adapted from the Eysenck Personality Questionnaire – Revised [8]. Note that the format of these items was modified by the first author 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 many other measures which are fully or partially overlapping, including abbreviated versions such as the 24 and 36 item Questionnaire Big Six scales [22], the 50 item IPIP scales corresponding to the Big Five factor markers [10], and the 20 item "mini- IPIP" scales [7].

The 1,034 items from these eight targeted 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 [20], a variant of matrix sampling procedures discussed by Lord [16]. This method produces data which contain "massive missingness" by design [19]. This missingness qualifies for classification as missing completely at random ["MCAR", 13] and it is further described as massively missing because the mean level of missingness by participant was 85.7% and 85.9% respectively for the Replication and Confirmatory Samples. The personality items were presented to participants in random order, and participants responded to as many items as they wished. Participants were encouraged to complete approximately 100 items but were able to complete up to 330.

Table 1 describes the size and density of personality data in all three samples. Variability in the number of responses was generally consistent across all three samples. The mean number of responses per participant ranged from 84.7 to 86.1 (SDs ranged from 57.2 to 58.7). Given this similarity in density, differences in the number of responses per item and in the number of pairwise responses were largely reflective of differences in the number of participants.

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 persistent 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.

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 [17], the data files are named: 'sapaTempData696items26jul2014thru22dec2015.RData' and 'sapaTempData696items22dec2015thru07feb2017.RData'.

The data files are named to indicate the domain (temperament), the number of items included (696), and the time period over which the data were collected (26 jul 2014 through 22 dec 2015 and 22 dec 2015 through 07 feb 2017). The files can be found at: https://doi.org/10.7910/DVN/GU70EV (Replication Sample) and https://doi.org/10.7910/DVN/TZJGAT (Confirmatory Sample).

The two data files each include four objects. The most pertinent of these is the raw data object ('sapaTempData696items26jul2014thru22dec2015' or 'sapaTempData696items22dec2015thru07feb2017'). The remaining three objects, which are identical in each data file, 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 have also been posted in a csv format at the same locations listed above. Each of the objects described above has been saved as a separate csv file, with the exception of the ItemLists because this is not easily reorganized into a spreadsheet format and could be easily re-created from the superKey csv file.

Data type

Self-report, cross-sectional survey data from 103,205 participants (54,855 participants from the Replication Sample and 48,350 from the Confirmatory Sample).

Format names and versions

The data are stored as separate sets of rdata files (approximately 10 MB each). Each file includes the four objects described above: the main data object, either 'sapaTempData696items26jul2014thru22dec2015.rdata' for the Replication Sample or 'sapaTempData696items 22dec2015thru07feb2016.rdata' for the Confirmatory Sample, 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, Lorien Elleman, Jason French, Elina Zaonegina, Zara Wright, and Sarah Russin 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 (https://dataverse.harvard.edu), within a separate dataverse for datasets relating to the SAPA-Project (https://dataverse.harvard.edu/dataverse/SAPA-Project). The Replication Sample is located at: https://doi.org/10.7910/DVN/GU70EV and the Confirmatory Sample is located at https://doi.org/10.7910/DVN/TZJGAT.

Publication date

The datasets were published on February 25, 2017.

(4) Reuse potential

The data are well-suited for many types of structural, correlational, and network 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 [18] 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.

Acknowledgements

The authors would like to acknowledge Joshua Wilt and Jason French for their contributions on the SAPA-Project.

Competing Interests

The authors have no competing interests to declare.

Author Contribution

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 third 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.

Ellen Roney, is a Research Assistant at Northwestern University's Feinberg School of Medicine in the Department of Medical Social Sciences. She contributed substantially to the preparation of this manuscript, including the production of the tables and figures, editing text, and preparing the data to be shared in the Dataverse.

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

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: https://doi.org/10.1016/j.paid.2006.10.027
2. **Condon, D M** 2014 An organizational framework for the psychological individual differences: Integrating the affective, cognitive, and conative domains. *PhD thesis*, Northwestern University, Evanston, IL.
3. **Condon, D M** (under review). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model.
4. **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: https://doi.org/10.5334/jopd.al
5. **DeYoung, C G** 2010 "Toward a Theory of the Big Five". *Psychological Inquiry*, 21(1): 26–33. DOI: https://doi.org/10.1080/10478401003648674
6. **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. PMid: 17983306. DOI: https://doi.org/10.1037/0022-3514.93.5.880
7. **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: https://doi.org/10.1037/1040-3590.18.2.192
8. **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: https://doi.org/10.1007/978-1-4613-2413-3
9. **Goldberg, L R** 1992 "The development of markers for the Big-Five factor structure". *Psychological Assessment*, 4(1): 26–42. DOI: https://doi.org/10.1037/1040-3590.4.1.26
10. **Goldberg, L R** 1999 "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor Models". In Mervielde, I, Deary, I, De Fruyt, F and Ostendorf, F (Eds.), *Personality and Individual Differences*, Tilburg University Press, The Netherlands, pp. 7–28.
11. **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].
12. **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: https://doi.org/10.1016/j.jrp.2005.08.007
13. **Graham, J W** 2009 "Missing Data Analysis: Making It Work in the Real World". *Annual Review of Psychology*, 60(1): 549–576. DOI: https://doi.org/10.1146/annurev.psych.58.110405.085530
14. **International Organization for Standardization** 2013 Codes for the representation of names of countries and their subdivisions. Part 1: Country codes. Geneva (Switzerland): The Organization. (ISO 3166-1: 2013).
15. **John, O P** and **Srivastava, S** 1999 "The Big Five trait taxonomy: History, measurement, and theoretical perspectives". In Pervin, L and John, O P (Eds.), *Handbook of personality: Theory and research*, Guilford Press, New York, pp. 102–138.
16. **Lord, F M** 1955 "Sampling fluctuations resulting from the sampling of test items". *Psychometrika*, 20(1): 1–22. DOI: https://doi.org/10.1007/BF02288956
17. **R Core Team** 2015 *"R: A Language and Environment for Statistical Computing"*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
18. **Revelle, W** 2016 *"psych: Procedures for psychological, psychometric, and personality research"*. Northwestern University, Evanston, Illinois, R package version 1.6.12.
19. **Revelle, W** and **Brown, A** 2013 "Standard errors for SAPA correlations". In *Society for Multivariate Experimental Psychology*, St. Petersburg, FL, pp. 1–10.
20. **Revelle, W, Condon, D M, Wilt, J, French, J A, Brown, A** and **Elleman, L G** 2016 Web and phone based data collection using planned missing designs. In: Fielding, N G, Lee, R M and Blank, G (Eds.), *Handbook of Online Research Methods*. Thousand Oaks, CA: Sage Publications.
21. **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: https://doi.org/10.4135/9781849200479.n13
22. **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: https://doi.org/10.1037/a0024165

How to cite this article: Condon, D M, Roney, E and Revelle, W 2017 A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. Journal of Open Psychology Data, 5: 3, DOI: https://doi.org/10.5334/jopd.32

Published: 09 June 2017

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

OPEN ACCESS