c-VEDA dataset main page: https://cveda-project.org/dataset/
Release date: 2018-11-21
DOI: 10.25720/veda-c09w

This early release has been obtained by manually processing the raw CSV files obtained from the Delosis server, according to the SOP detailed below. One of the main objectives of this effort was to assess the quality of the collected data, detect errors such as subject identifier misassignments and fix these errors.

Psytools files can be downloaded via SFTP:
`sftp://cveda.nimhans.ac.in/data/0.9/psytools/`

Errors and caveats

Among other issues, we have found issues in the pseudonymization of a dozen participants. We recommend you wait for release 1.1.

Version 0.9 (21-Nov-2018)

c-VEDA Data Quality Checking (QC) Standard Operation Procedure (SOP)
- Psytools -

* This SOP is specific for c-VEDA psytools data release V0.9 (21-Nov-2018). The procedure will be improved with time. Please check the release notes specific to the release version.

Step 1: Download the csv files from sftp://veda.nimhans.ac.in/data/processed/psytools/.

Step 2: Open your .csv file and save it as Excel Workbook locally

- One file per questionnaire that includes the data from all subjects
- Make sure to back up the data and always work with copies

Step 3: Scan data for the following:

1) Column labels
   a) Are all items from the questionnaire included?
   b) Are the labelling of the items correct?
2) Strange values (most of these will be picked up with the help of R, but any obvious oddities could be identified here as well)
   a) Are the scaling of the items correct (e.g., values within the correct range, check the questionnaire);
   b) Any strange values.

Step 4: Check for duplicates

This can be done in excel using the following commands (after selecting the entire column A):

> *Select entire column A (usercode) > Home > Conditional formatting > Highlight cell rules > Duplicate values > (Style: Classic; "Format only unique or duplicate values"; "duplicate" values in the selected range; Format with "light red fill with dark red text") > Ok*

Duplicates need to be resolved based on the following:

1) If all iterations under the same PSC1 code have the same age band:
   a) Leave the one with the highest iteration number and delete the others;
   b) IMPORTANT! In the case of there being multiple iterations, but the row with iteration 2 having incomplete data, then remove the row with iteration 2
2) If there are different age bands recorded for duplicates of the same PSC1 code:
   a) Check recruitment file to verify which age band is the correct one;
   b) remove all of the duplicates with the wrong age band and leave only the one duplicate with the highest iteration number
3) If both age bands have complete or incomplete datasets – recruitment team at the site to be contacted to identify the correct age band/dataset

Step 5: Save the data as an excel file

Step 6: Import data into Rcmdr or any statistical program (e.g., SPSS)

Step 7: Compute summary scores & Data modification (refer to Appendix 1)

Step 8: Data is to be analysed to obtain the following QC measures:

1) Compute total score
2) Total number of subjects
3) Distribution based on – site, age band, gender
4) Strange values in dataset – to be listed along with the PSC1 code
5) Refused values in dataset (R, NR) – to be listed along with the PSC1 code
6) Missing values in dataset – PSC1 codes to be listed with details of missing values
    a. Check frequency of missing values under each variable
    b. Open to the data file in excel – filter individual variables by "blanks" – note down the PSC1 codes for each variable – under the specific site columns
    c. Make sure there are as many PSC1 codes as the 'missing values' reported in R
7) Summaries – Frequencies (%) for count data; Mean (SD) for score data
8) Outliers – to be listed
    a. Outliers are values that are less or more than 2 SD of the mean value.
    b. To be calculated site and age band wise.
    c. Suppose the mean value for the whole sample for "TotalDifficulty" is 10, and the SD is 2.5.
    d. Any value that is less or more than 2 times the SD, i.e. 2*2.5=5, is an outlier. So any value <5 (i.e. 10 – 2*SD) or >15 (i.e. 10+2*SD) would be an outlier.
    e. In R you can identify outliers using the following steps.
    Open Data > Active data set > Subset active data set… > Select "Include all variables">From the variable list select the variable for which you want to get the outliers, i.e. in this case it would be "TotalDifficulty">In 'Subset expression' – write the condition for calling a value a outlier – "TotalDifficulty<5 | TotalDifficulty>15">Name for new dataset: can be called anything - "TotalDifficulty.Outliers">Press "Ok"
    f. After this one can view the outlier PSC1 codes by clicking on "view dataset". The dataset here is the one we have just created "TotalDifficulty.Outliers". The PSC1 codes from here are to be copied on the slide.
    g. Same procedure to be carried out for ALL summary variables. Please note: If the Outlier datasets created show "0 rows" – it means that there are no outliers for that variable.
9) Group comparisons for – Gender, Site and Age effects

Step 9: Convert PSC1 codes to PSC2 codes.

Appendix 1 Generation of Summary Variables & data modification rules

Questionnaire	Data modification	Dimensional variable	Computation
Adolescents Attachment Questionnaire (AAQ)	Reverse scoring for Availability and Goal Corrected Partnership	Anger Distress	Sum of all variables
		Availability	Sum of all variables
		Goal Corrected Partnership	Sum of all variables
Adverse childhood experiences – International questionnaire (ACE-IQ)		For binary score	
		Emotional abuse	A1 OR A2 >0
		Physical abuse	A3 OR A4 >0
		Sexual abuse	A5 OR A6 OR A7 OR A8 >0
		Violence	F6 OR F7 OR F8 >0
		HH member SUD	F1 >0
		HH member MI	F2 >0
		HH member jailed	F3 >0
		Parental loss	F4 OR F5 >0
		Emotional neglect	P1 OR P2 >0
		Physical neglect	P3 OR P4 OR P5 >0
		Bullying	V1 >0
		Community violence	V4 OR V5 OR V6 >0
		Collective violence	V7 OR V8 OR V9 OR V10 >0
		For Frequency score	
		Emotional abuse	A1 OR A2 =3
		Physical abuse	A3 OR A4 =3
		Sexual abuse	A5 OR A6 OR A7 OR A8 =1/2/3
		Violence	F6=3 OR F7=2/3 OR F8=2/3
		HH member SUD	F1 =1
		HH member MI	F2 =1
		HH member jailed	F3 =1
		Parental loss	F4 OR F5 =1
		Emotional neglect	P1 OR P2 =0/1
		Physical neglect	P3 OR P4 OR P5 =3
		Bullying	V1 =3
		Community violence	V4 OR V5 OR V6 =3
		Collective violence	V7 OR V8 OR V9 OR V10 =1/2/3
		Adversity.Binary	Sum of all binary variables
		Adversity.Frequency	Sum of all frequency variables
		Addendum score	Sum of Ad1 to Ad5
		CRIES.Abuse	Sum of CRIES_5_1 to
		CRIES.Bullying	Sum of CRIES_6_1 to
		CRIES.Collective Violence	Sum of CRIES_8_1 to
		CRIES.Community	Sum of CRIES_7_1 to
		CRIES.Family	Sum of CRIES_4_1 to
		CRIES.Neglect	Sum of CRIES_3_1 to
Alabama parenting questionnaire	Replace all “R”, “NR”, “NA” response entries with <empty space>	Involvement (Child form:	1+4+7+9+11+14+15+20+23+26
		Positive parenting	2+5+13+16+18+27
		Poor monitoring	6+10+17+19+21+24+28+29+30

(APQ) - Child, Parent		Inconsistent discipline	3+8+12+22+25+31
		Corporal punishment	33+35+39
Alcohol, smoking and substance involvement screening test (ASSIST)	Coding change	Specific substance involvement score	Sum question 3, 5, 6, 7, 8, 9 for
	Q3		a: prescription
	0 = 0		b: tobacco (range 0 – 31)
	10 = 2		c: alcohol (range 0 – 39)
	20 = 3		d: cannabis (range 0 – 39)
	30 = 4		e: inhalants (range 0 – 39)
	40 = 6		f: sleeping pills
	For Q6:		g: opioids
	0 = 0		h: amphetamine-type
	10 = 4		i: cocaine (range 0 – 39)
	20 = 5		j: hallucinogens (range 0 – 39) i
	30 = 6		K specify:
	40 = 7		
	For Q7:	Low risk	Alcohol: 0-10
	0 = 0		Others: 0-3
	10 = 5	Moderate risk	Alcohol: 11-26
	20 = 6		Others: 4-26
	30 = 7	High risk	Alcohol: >/=27
	40 = 8		Others: >/=27
Big Five Inventory BFI)	1. Replace all "R"; *2.* Reverse score items: 2, 6, 8, 9, 12, 18, 21, 23, 24, 27, 31, 34, 35, 37, 41, 43 (1-5, 2-4, 3-3, 4-2, 5-1) "NR", "NA" response entries with <empty space>	Extraversion	1+6R+11+16+21R+26+31R+36
		Agreeableness	2R, 7, 12R, 17, 22, 27R, 32, 37R,
		Conscientiousness	3+8R+13+18R+23R+28+33+38+
		Neuroticism	4+9R+14+19+24R+29+34R+39
		Openness	5+10+15+20+25+30+35R+40+41R+44
Parental bonding instrument (PBI)		Care (mother & father	1+2+4+5+6+11+12+14+16+17+1
		Overprotection (mother &	3+7+8+9+10+13+15+19+20+21+
Puberty development scale (PDS)	*For boys* Consider variables 'Hair growth', 'Voice change', 'Facial hair' Calculate SUM	*For boys*	
		Prepubertal	3
		Early	04-May
		Mid	06-Aug
		Late	09-Nov
		Post	12
	For girls: - Consider variables 'Hair growth', 'Breast growth'; - Calculate SUM + Consider 'Menses started?'	*For girls*	
		Prepubertal	2
		Early	3
		Mid	4
		Late	</=7
		Post	8

Indian Family Violence and Control Scale (IFVCS)		Control score	Sum of all CONTROL variables
		Physical Abuse	Sum of all PHYSICAL variables
		Psychological Abuse	Sum of all PSYCH variables
		Sexual Abuse	Sum of all SEXUAL variables
School climate questionnaire (SCQ)	1. Replace all "R", "NR", "NA" response entries with <empty space>; 2. Reverse score items: 5, 9, 13, 17 (1-4, 2-3, 3-2, 4-1)	Safety & Order	1+5R+9R+13R+17R
		Support & Acceptance	2+6+10+14+18
		Equity & Fairness	3+7+11+15+19
		Encouraging, autonomy & cooperativeness	4+8+12+16+20+21
Strengths and Difficulties Questionnaire (SDQ) - Child, Adult, Parent	1. Replace all "R", "NR", "NA" response entries with <empty space>; 2. Reverse score items: 7, 11, 14, 21, 25 (1=3, 2=2, 3=1)	Emotional symptom	3+8+13+16+24
		Conduct problem	5 + 7R + 12 + 18 + 22
		Hyperactivity/inattention	2 + 10 + 15 + 21R + 25R
		Peer problem	6 + 11R + 14R + 19 + 23
		Prosocial behaviour	1 + 4 + 9 + 17 + 20
		Externalizing difficulties	Conduct problems +
		Internalizing difficulties	Peer problems + Emotional
		Total difficulty	Externalizing + Internalizing