**Flinders University**
Digital Psychiatry
Research Team

# Transformation of the OWC dataset: Data Cleaning and Augmentation Procedures

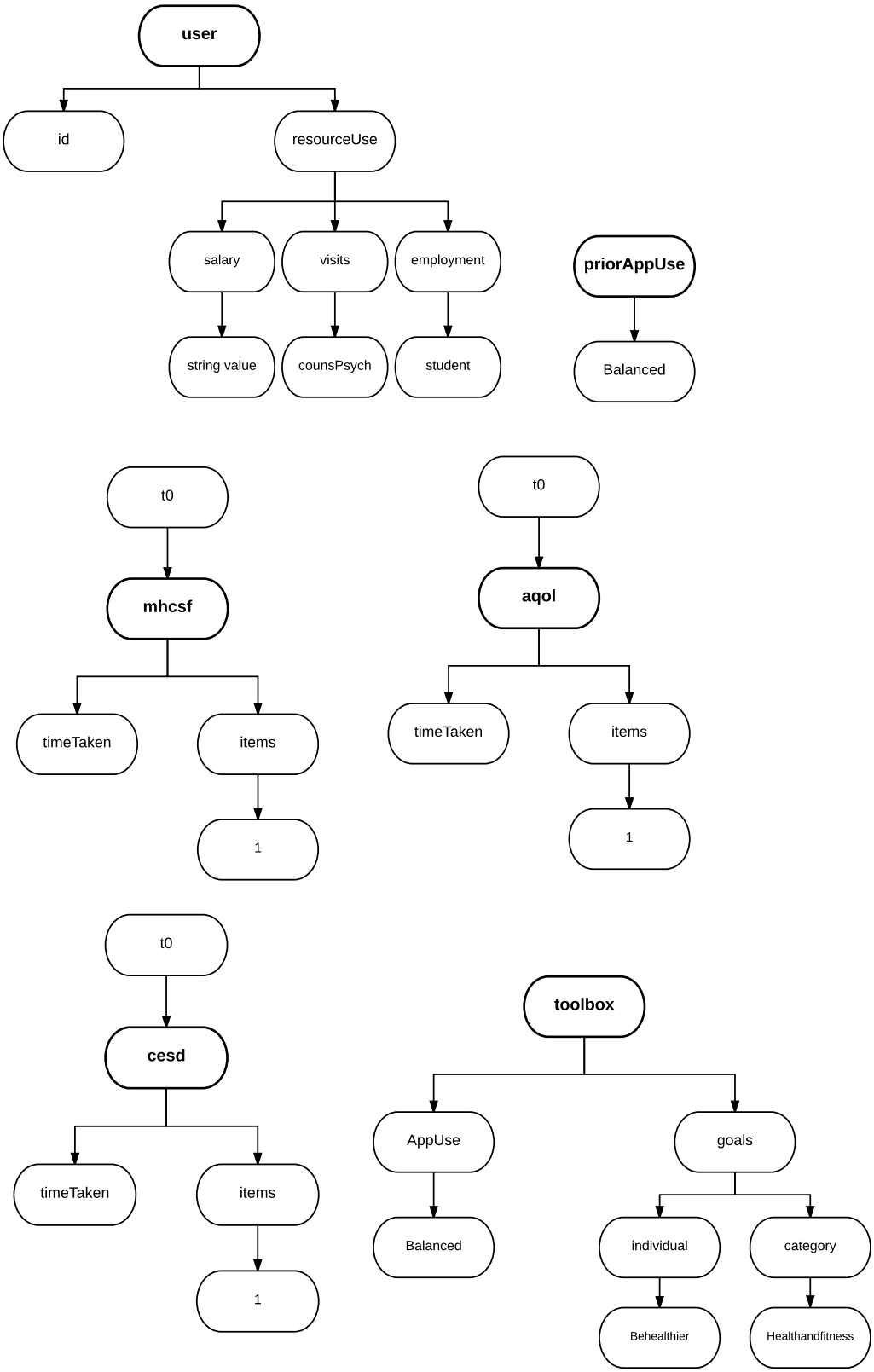
## Removal of 'Test' Data

All 'test' user accounts were removed from the dataset. These can be described as accounts which researchers on the project used to test functioning of the study platform and study measures before and during the OWC study trial. Test accounts were identified and removed based on the following:

1. All those accounts which were created or used to complete study measures before 9am on 31st October 2014. This date represents the beginning of active recruitment for the OWC trial, therefore all accounts created and used before this date could be identified as pre-trial internal test accounts.
2. Those accounts which were created after 9am on 31st October 2014, however could be identified as test accounts used by the project researchers based on the sign up name and/or email address. The unique 3- or 4- number study platform ids for these are listed below:
    - 929
    - 930
    - 949
    - 966
    - 973
    - 1117
    - 1247
    - 1299
    - 1310
    - 1474
    - 1477
    - 1489

## Naming of variables

Variable names were created based on the broader variable type or data 'category', then narrowing down to the variable specifics. We identified 6 data categories, namely 'user', 'priorAppUse', 'aqol', 'cesd', 'mhcsf', and 'toolbox'. Examples of the naming convention are as follows:




A codebook of all variables has also been provided.

**Salary coding**

Participants were asked to provide information on their fortnightly salary. This

information was collected as open text, retained in the variable *user.resourceUse.salary.stringValue*. In order to facilitate analysis, an additional variable with a numeric representation (in A$ per fortnight) of this data was created *users.resourceUse.salary*. For the recoding of the text into a numeric representation, the following rules were applied:

- simple numerical values were copied (e.g. “$2000 per fortnight” = 2000)
- in the case of numerical values with additional information or quantifiers, additional information was disregarded (e.g. “around $2000” = 2000, “$1200 after tax” = 1200)
- where users provided a range, the mean was taken (e.g. “$100-$300” = 200)

information without any specific figures was disregarded (e.g. “depends on hours” = MISSING)

**Description of MHC-SF and sub scales**

The Mental Health Continuum short form (MHC-SF) is derived from the long form (MHC-LF), which consists of 40 items (or questions) measuring 3 distinct facets of well-being (emotional, psychological and social). The MHC-SF consists of 14 items derived from the MHC-LF. These 14 items are divided up into the 3 wellbeing facets as follows:

- 3 items represent emotional well-being,
- 6 items represent psychological well-being, and
- 5 items represent social well-being.

Along with these facets the MHC-SF also assesses and categorises levels of positive mental health; into *flourishing*, *moderate* or *languishing* mental health. The MHC-SF has shown excellent internal consistency (> .80) and discriminant validity in adolescents (ages12-18) and adults (Keyes, C. L. M. 2009. Atlanta: Brief description of the mental health continuum short form (MHC-SF). Available: http://www.sociology.emory.edu/ckeyes/. [On–line, retrieved 18 August 2016]).

To ascertain the integrity of the MHC-SF data for each individual and measurement time, the number of missing items was computed across the questionnaire. The following SPSS syntax was used (note that variables created in this syntax were temporary and not retained in the final data set, X refers to the corresponding measurement time [0-3]):

```
COUNT tX.mhcsf.missing = tX.mhcsf.items.1 TO tX.mhcsf.items.14 (MISSING) .
EXECUTE .
```

For all time points, there was only complete data for each individual or completely missing, i.e. there were no cases in the dataset where individuals only completed a subset of items on this measure. For this reason, MHC-SF scores were calculated for each individual where data was available, all scores are subsequently based on complete data and there was no imputation of data.

To calculate the MHC-SF total score, a sum of all item responses was calculated using the following SPSS syntax:

```
* calculate totals .
```

```
COMPUTE tX.mhcsf.score = SUM(tX.mhcsf.items.1 TO tX.mhcsf.items.14) .
COMPUTE tX.mhcsf.ewb = MEAN(tX.mhcsf.items.1 TO tX.mhcsf.items.3) .
COMPUTE tX.mhcsf.swb = MEAN(tX.mhcsf.items.4 TO tX.mhcsf.items.8) .
COMPUTE tX.mhcsf.pwb = MEAN(tX.mhcsf.items.9 TO tX.mhcsf.items.14) .
EXECUTE .
```

```
EXECUTE .
```

Diagnoses were computed according to the recommendation outlined in Keyes, C. L. M. 2014. Overview of the Mental Health Continuum Short Form (MHC-SF). Emory University. http://www.bttop.org/sites/default/files/public/MHC-SF%20Brief%20Introduction%209.22.2014.pdf , which calculates mean (as opposed to total) scores.

The following SPSS syntax was used for each timepoint (note that only 2 and 3 category diagnosis variables were retained).

```
COUNT hiaff=tX.mhcsf.items.1 tX.mhcsf.items.2 tX.mhcsf.items.3(4,5).
COUNT loaff=tX.mhcsf.items.1 tX.mhcsf.items.2 tX.mhcsf.items.3(0,1).
COUNT hifunc=tX.mhcsf.items.4 tX.mhcsf.items.5 tX.mhcsf.items.6
tX.mhcsf.items.7 tX.mhcsf.items.8 tX.mhcsf.items.9 tX.mhcsf.items.10
tX.mhcsf.items.11 tX.mhcsf.items.12 tX.mhcsf.items.13 tX.mhcsf.items.14(4,5).
COUNT lofunc=tX.mhcsf.items.4 tX.mhcsf.items.5 tX.mhcsf.items.6
tX.mhcsf.items.7 tX.mhcsf.items.8 tX.mhcsf.items.9 tX.mhcsf.items.10
tX.mhcsf.items.11 tX.mhcsf.items.12 tX.mhcsf.items.13 tX.mhcsf.items.14(0,1).
EXECUTE .
RECODE hiaff (1,2,3=1) (else=0) into hiaffect.
RECODE hifunc (6,7,8,9,10,11=1) (else=0) into hifunct.
RECODE loaff (1,2,3=1) (else=0) into loaffect.
RECODE lofunc (6,7,8,9,10,11=1) (else=0) into lofunct.
EXECUTE .

IF ((hiaff>0) OR (hifunc>0) OR (loaff>0) OR (lofunc>0)) tX.mhcsf.dx=1.
IF hiaffect=1 and hifunct=1 tX.mhcsf.dx=2.
IF loaffect=1 and lofunct=1 tX.mhcsf.dx=0.
EXECUTE .
RECODE tX.mhcsf.dx (0=0) (2=2) (1=1) into tX.mhcsf.dx3.
RECODE tX.mhcsf.dx3 (2=1) (1=0) (0=0) into tX.mhcsf.dx2.
VARIABLE LABELS tX.mhcsf.dx3 'TX MHC-SF Three Category Diagnosis of Positive
Mental Health'.
VALUE LABELS tX.mhcsf.dx3 0 'Languishing' 1 'Moderate' 2 'Flourishing'.
```

```
VARIABLE LABELS tX.mhcsf.dx2 'TX MHC-SF Two Category Diagnosis of Positive
Mental Health'.
VALUE LABELS tX.mhcsf.dx2 0 'Not Flourishing' 1 'Flourishing'.
EXECUTE .

DELETE VARIABLES hiaff loaff hifunc lofunc hiaffect loaffect hifunct lofunct
tX.mhcsf.dx .
```

**Description of Aqol-4D**

The Assessment of Quality of Life (AQoL) questionnaires are health-related quality of life instruments, designed for use in economic evaluation studies (http://www.aqol.com.au/index.php/what-is-aqol). The AQoL-4D instrument consists of 12 items (or questions) designed to assess 4 distinct dimensions: Independent Living, Mental Health, Relationships, and Senses (http://www.aqol.com.au/index.php/choice-of-aqol-instrument).

The following syntax was used to derive the AQoL-4D dimension and utility scores (http://www.aqol.com.au/index.php/scoring-algorithms?id=82):

```
IF (tX.aqol.items.1 = 1) dvQ1 = 0.000.
IF (tX.aqol.items.1 = 2) dvQ1 = 0.154.
IF (tX.aqol.items.1 = 3) dvQ1 = 0.403.
IF (tX.aqol.items.1 = 4) dvQ1 = 1.000.

IF (tX.aqol.items.2 = 1) dvQ2 = 0.000.
IF (tX.aqol.items.2 = 2) dvQ2 = 0.244.
IF (tX.aqol.items.2 = 3) dvQ2 = 0.343.
IF (tX.aqol.items.2 = 4) dvQ2 = 1.000.

IF (tX.aqol.items.3 = 1) dvQ3 = 0.000.
IF (tX.aqol.items.3 = 2) dvQ3 = 0.326.
IF (tX.aqol.items.3 = 3) dvQ3 = 0.415.
IF (tX.aqol.items.3 = 4) dvQ3 = 1.000.

IF (tX.aqol.items.4 = 1) dvQ4 = 0.000.
IF (tX.aqol.items.4 = 2) dvQ4 = 0.169.
IF (tX.aqol.items.4 = 3) dvQ4 = 0.396.
IF (tX.aqol.items.4 = 4) dvQ4 = 1.000.

IF (tX.aqol.items.5 = 1) dvQ5 = 0.000.
IF (tX.aqol.items.5 = 2) dvQ5 = 0.095.
IF (tX.aqol.items.5 = 3) dvQ5 = 0.191.
IF (tX.aqol.items.5 = 4) dvQ5 = 1.000.

IF (tX.aqol.items.6 = 1) dvQ6 = 0.000.
IF (tX.aqol.items.6 = 2) dvQ6 = 0.147.
```

```
IF (tX.aqol.items.6 = 3) dvQ6 = 0.297.
IF (tX.aqol.items.6 = 4) dvQ6 = 1.000.

IF (tX.aqol.items.7 = 1) dvQ7 = 0.000.
IF (tX.aqol.items.7 = 2) dvQ7 = 0.145.
IF (tX.aqol.items.7 = 3) dvQ7 = 0.288.
IF (tX.aqol.items.7 = 4) dvQ7 = 1.000.

IF (tX.aqol.items.8 = 1) dvQ8 = 0.000.
IF (tX.aqol.items.8 = 2) dvQ8 = 0.253.
IF (tX.aqol.items.8 = 3) dvQ8 = 0.478.
IF (tX.aqol.items.8 = 4) dvQ8 = 1.000.

IF (tX.aqol.items.9 = 1) dvQ9 = 0.000.
IF (tX.aqol.items.9 = 2) dvQ9 = 0.219.
IF (tX.aqol.items.9 = 3) dvQ9 = 0.343.
IF (tX.aqol.items.9 = 4) dvQ9 = 1.000.

IF (tX.aqol.items.10 = 1) dvQ10 = 0.000.
IF (tX.aqol.items.10 = 2) dvQ10 = 0.107.
IF (tX.aqol.items.10 = 3) dvQ10 = 0.109.
IF (tX.aqol.items.10 = 4) dvQ10 = 1.000.

IF (tX.aqol.items.11 = 1) dvQ11 = 0.000.
IF (tX.aqol.items.11 = 2) dvQ11 = 0.141.
IF (tX.aqol.items.11 = 3) dvQ11 = 0.199.
IF (tX.aqol.items.11 = 4) dvQ11 = 1.000.

IF (tX.aqol.items.12 = 1) dvQ12 = 0.000.
IF (tX.aqol.items.12 = 2) dvQ12 = 0.104.
IF (tX.aqol.items.12 = 3) dvQ12 = 0.312.
IF (tX.aqol.items.12 = 4) dvQ12 = 1.000.
EXECUTE.

COMPUTE dvD1= (1.0989*(1-(1-0.6097*dvQ1)*(1-0.4641*dvQ2)*(1-0.5733*dvQ3))).
COMPUTE dvD2 = (1.0395*(1-(1-0.7023*dvQ4)*(1-0.6253*dvQ5)*(1-
0.6638*dvQ6))).
COMPUTE dvD3 = (1.6556*(1-(1-0.2476*dvQ7)*(1-0.2054*dvQ8)*(1-
0.3382*dvQ9))).
COMPUTE dvD4= (1.2920*(1-(1-0.1703*dvQ10)*(1-0.2554*dvQ11)*(1-
0.6347*dvQ12))).
EXECUTE .

COMPUTE tX.aqol.dimensions.inliv = 1-dvD1.
COMPUTE tX.aqol.dimensions.relat= 1-dvD2 .
COMPUTE tX.aqol.dimensions.sense = 1-dvD3 .
COMPUTE tX.aqol.dimensions.menth = 1-dvD4.
EXECUTE .

***INSTRUMENT SCORE
```

*** This model uses W = 1.04.

```
COMPUTE tX.aqol.utilitySc  = ((1.04* ((1-(0.841*dvD1))*  (1-(0.855*dvD2 ))*  (1-
(0.931*dvD3))*  (1-(0.997*dvD4)))) - 0.04).
EXECUTE.

VARIABLE LABELS tX.aqol.utilitySc 'AQoL4D Utility Score'
tX.aqol.dimensions.inliv 'TX AQOL Independent Living dimension score'
tX.aqol.dimensions.relat  'TX AQOL Relationships dimension score'
tX.aqol.dimensions.sense  'TX AQOL Senses dimension score'
tX.aqol.dimensions.menth  'TX AQOL Mental Health dimension score'.
EXECUTE.

DELETE VARIABLES dvQ1 dvQ2 dvQ3 dvQ4 dvQ5 dvQ6 dvQ7 dvQ8 dvQ9 dvQ10
dvQ11 dvQ12 dvD1 dvD2 dvD3 dvD4 .
EXECUTE.
```

**Description of CES-D**

The CES-D instrument is a 20-item self-report scale designed to measure depressive symptomology (such as restless sleep, poor appetite, and feeling lonely) in the general population. The possible range of total scores is zero to 60, with the highest scores indicating more depression symptoms. The CES-D also provides a cutoff score of 16, with those scoring higher than this considered at risk of clinical depression. (Radloff, L.S. (1977). The CES-D scale: A self-report depression scale for research in the general population. Applied Psychological Measurement, 1: 385-401)

To calculate the score for the CES-D, a sum of all items was calculated using the following syntax. These scores are all based on complete data, no imputation of missing items was necessary across the dataset:

```
COMPUTE cesd.score = SUM(cesd.items.1 TO cesd.items.20) .
EXECUTE .

FORMATS cesd.score (F8.0) .
VARIABLE LABELS cesd.score 'CES-D score' .
```

**Deidentification**

The following variables were removed from the dataset in the deidentification process:

user.firstName
user.lastName
user.email
user.mobile

Additionally, there was one user who identified as ‘genderqueer’ when asked to select their gender. Considered an outlier, this individual’s data was also removed from the dataset.