Barcelona crime longitudinal data quality

Alexandru Cernat

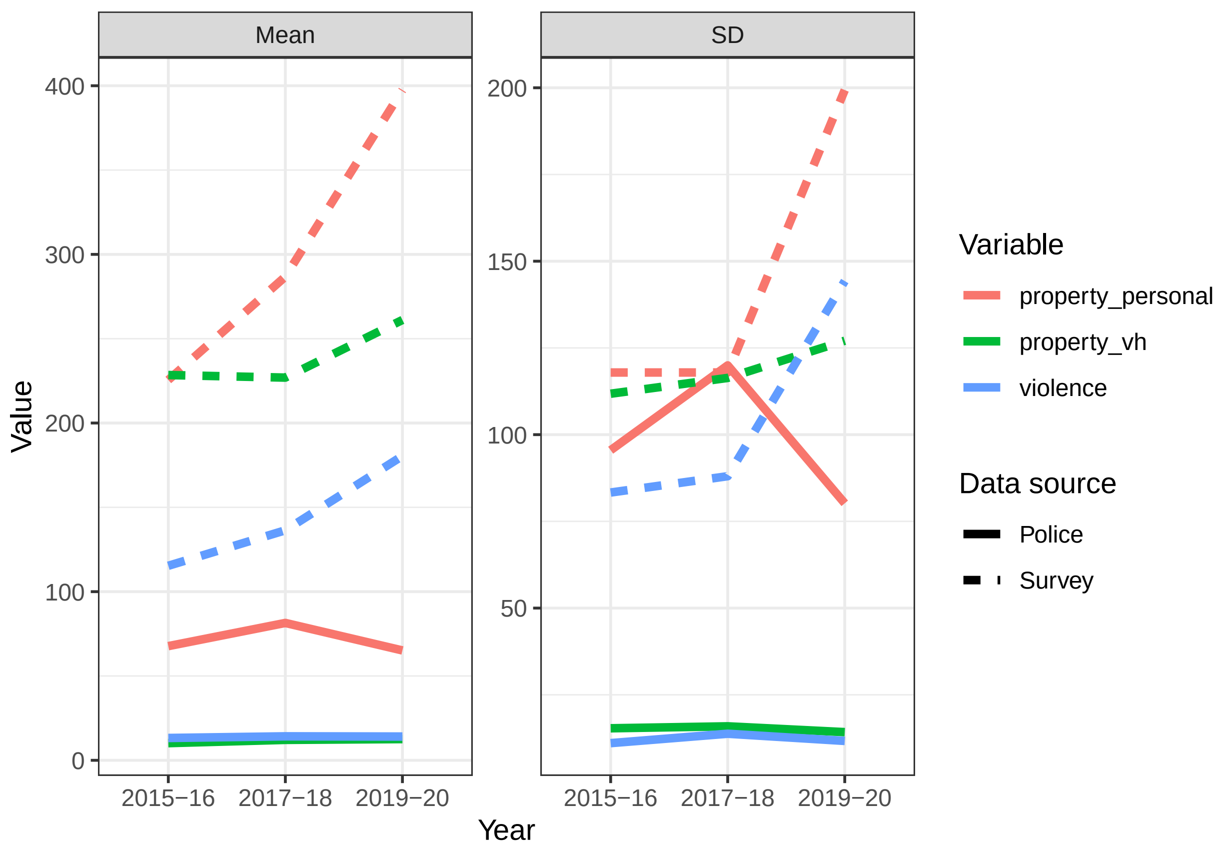
2021-03-11

Here I explore the Barcelona data that has three types of crimes: property vehicle, property personal and violence collected in surveys and official data over regions (76) and time (6 years). We group years by two in order to avoid having regions with 0s.

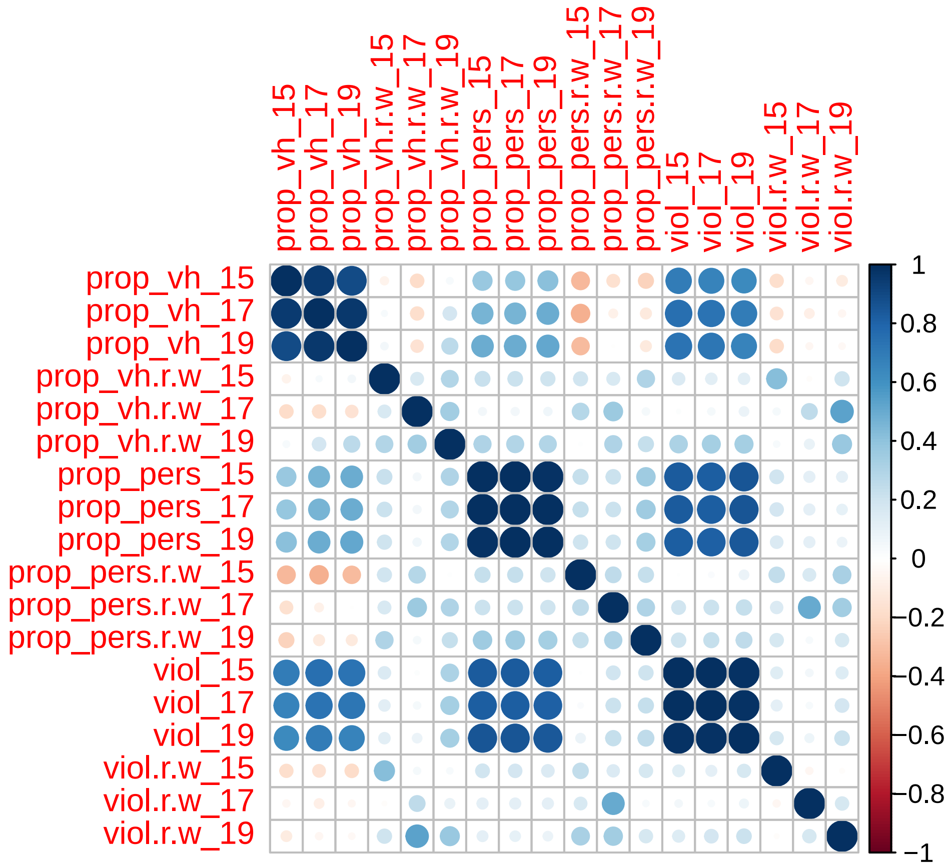
Here I concentrate on the weighted estimates from the survey (ending in "r.w") and the official data.

Descriptives

First some descriptives. Bellow we observe quite big differences by data source.

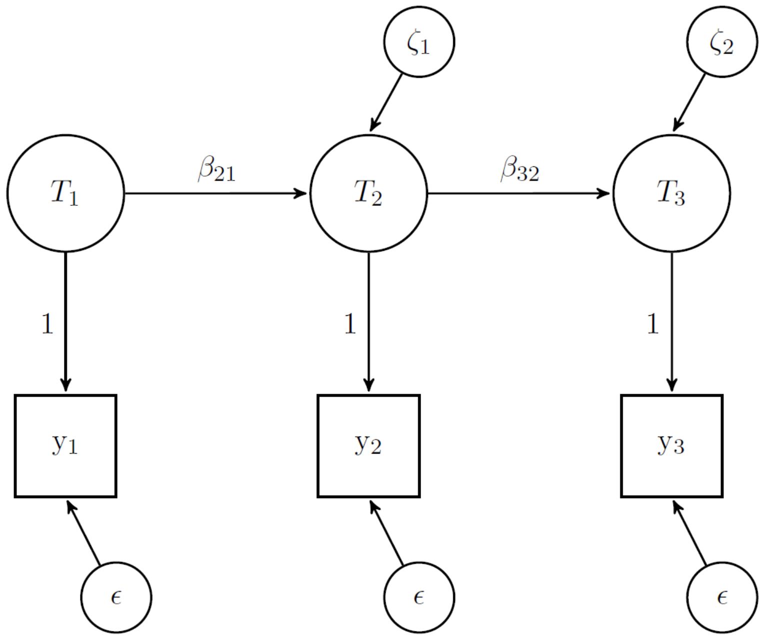


When we look at correlations we also see pretty striking patterns. First of all the consistency within measure is much higher for police data than survey data. Then, the relationship between of measures across data sources is very low. This could be problematic for any MTMM like modeling.



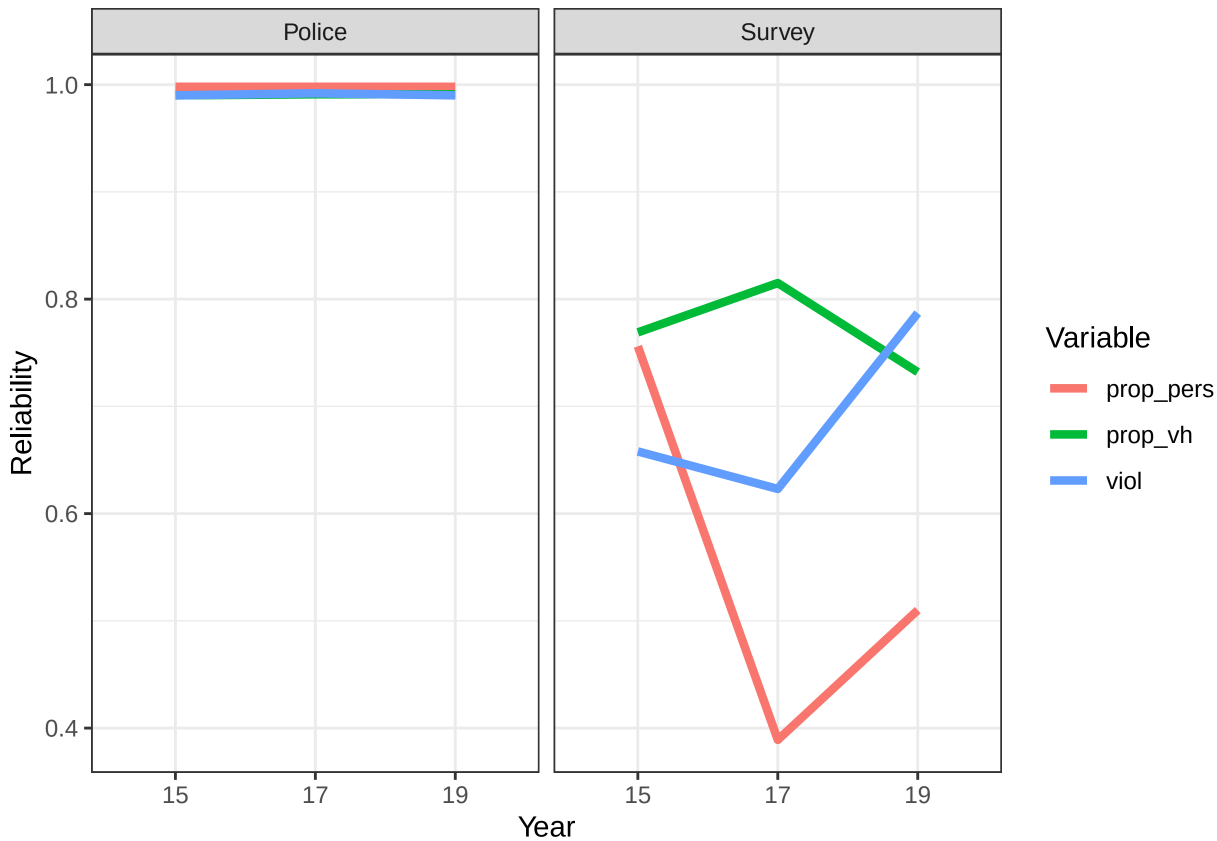
Quasi-simplex

A first way to look at the quality of the data is to use the quasi-simplex model. This assumes an auto-regressive model of true scores and estimates reliability by assuming equal variance of error over time (see for more details: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2014-09.pdf).



I estimate the models using `blavaan` which does in the background SEM using Stan. *I tried to use ML but it leads to negative variances (relative common occurrence for these models).*

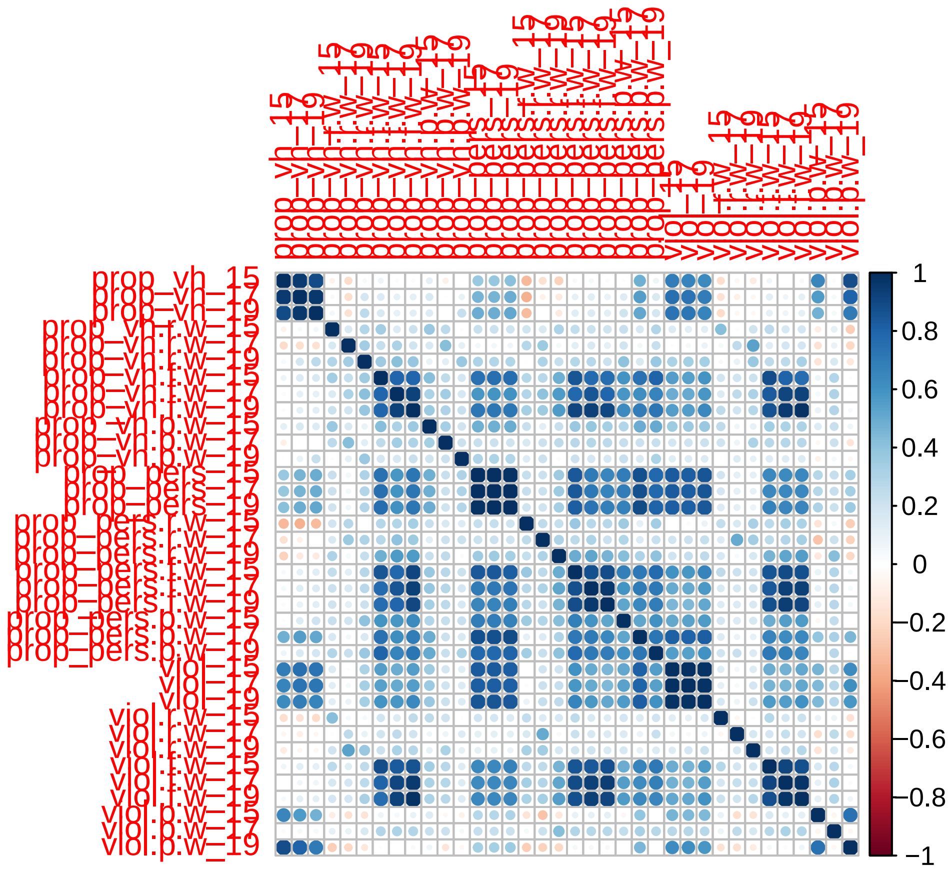
Here we plot the reliability by variable, data source and wave as estimated by quasi-simplex. Reliabilities are much lower for the survey data (as expected given the correlation matrices above).

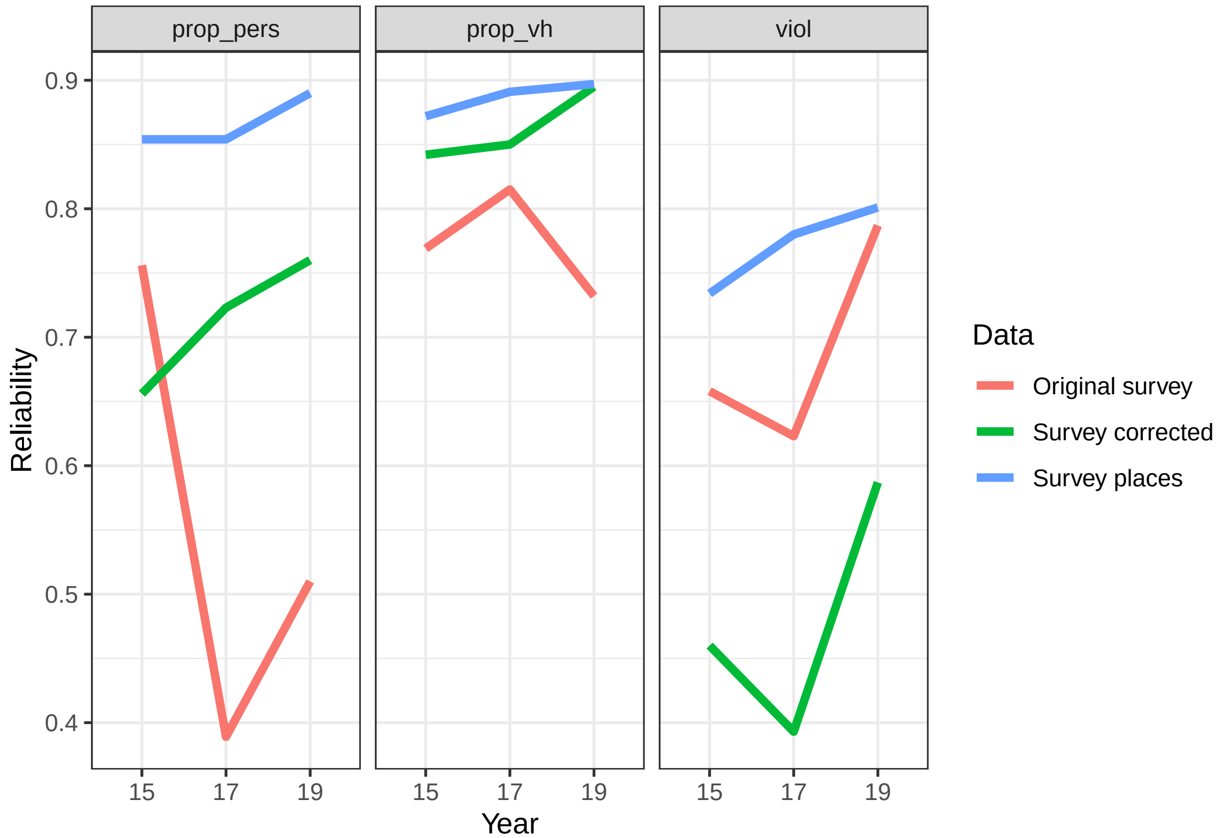


This is also obvious if we average reliability estimates.

group	var	reliability	reliability_source
Police	prop_pers	1.00	0.99
Police	prop_vh	0.99	0.99
Police	viol	0.99	0.99
Survey	prop_pers	0.55	0.67
Survey	prop_vh	0.77	0.67
Survey	viol	0.69	0.67

Given this and the point David made in a recent email let's have a look at the survey data about places where crimes happen (".f.w" - "Survey places") and looking only at crimes reported to the police (".p.w" - "Survey corrected").



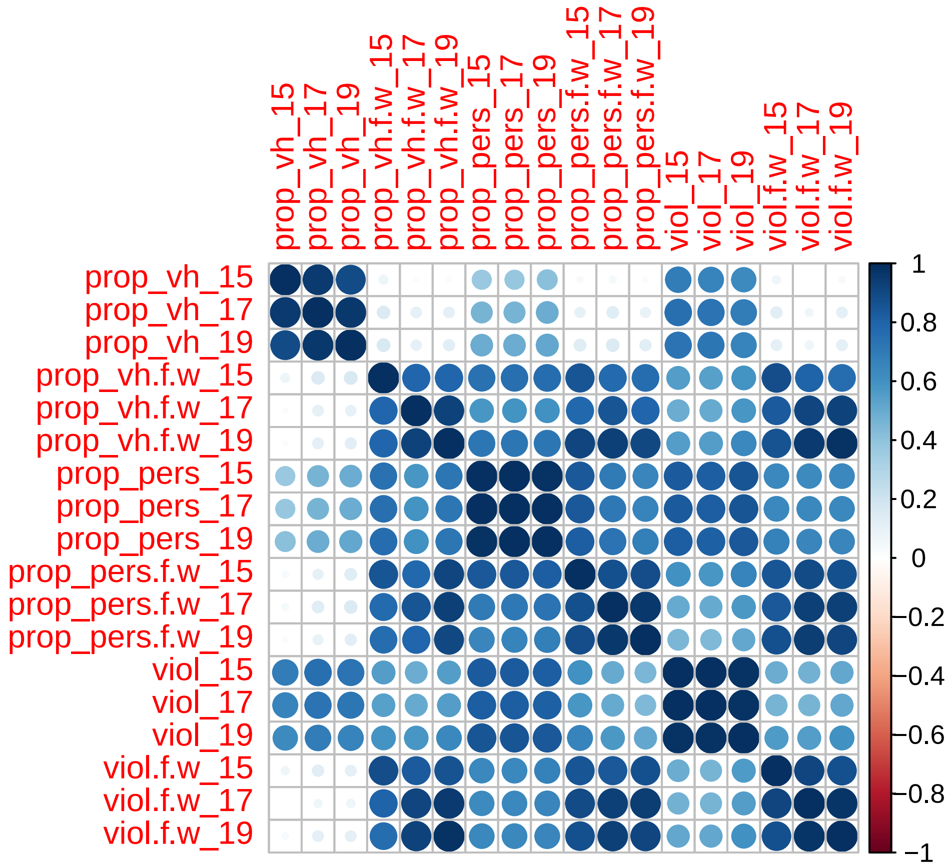


var	group	reliability	reliability_source
prop_pers	Original survey	0.55	0.67
prop_pers	Survey corrected	0.71	0.69
prop_pers	Survey places	0.87	0.84
prop_vh	Original survey	0.77	0.67
prop_vh	Survey corrected	0.86	0.69
prop_vh	Survey places	0.89	0.84
viol	Original survey	0.69	0.67
viol	Survey corrected	0.48	0.69
viol	Survey places	0.77	0.84

Overall seems that the ".p.w" measures have better quality so we'll use them in the next step.

Longitudinal variance decomposition

Next we will expand the quasi-simplex. We will make a measurement model at each wave that includes the measure of interest from the police and the corrected survey data in addition to include the simplex change in time. The statistic of interest here would be the standardized loading on trait which can be seen as "validity".



Property vehicle

```
## blavaan (0.3-15) results of 2000 samples after 8000 adapt/burnin iterations
##
##   Number of observations                            73
##
##   Number of missing patterns                         1
##
##   Statistic                                 MargLogLik         PPP
##   Value                                       -645.804       0.000
##
## Latent Variables:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t1 =~
##     prop_vh_15        1.000                               0.627    0.959
##     prop_vh.f.w_15    1.000                               0.627    0.258
##   t2 =~
##     prop_vh_17        1.000                               0.685    0.994
##     prop_vh.f.w_17    1.000                               0.685    0.254
##   t3 =~
##     prop_vh_19        1.000                               0.680    0.974
##     prop_vh.f.w_19    1.000                               0.680    0.257
##      Rhat    Prior
##
##        NA
```

```
##        NA
##
##        NA
##        NA
##
##        NA
##        NA
##
## Regressions:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t2 ~
##     t1                1.069    0.069    0.936    1.204    0.978    0.978
##   t3 ~
##     t2                0.970    0.041     0.89    1.052    0.978    0.978
##      Rhat    Prior
##
##     1.000    normal(0,10)
##
##     1.000    normal(0,10)
##
## Intercepts:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .prop_vh_15        2.112    0.077    1.961    2.264    2.112    3.233
##    .prop_vh.f.w_15    7.445    0.282    6.887    8.001    7.445    3.064
##    .prop_vh_17        2.267    0.081    2.106    2.426    2.267    3.290
##    .prop_vh.f.w_17    7.662    0.319    7.032    8.292    7.662    2.845
##    .prop_vh_19        2.319    0.082    2.157    2.481    2.319    3.322
##    .prop_vh.f.w_19    7.622    0.307    7.018     8.22    7.622    2.880
##     t1                0.000                               0.000    0.000
##    .t2                0.000                               0.000    0.000
##    .t3                0.000                               0.000    0.000
##      Rhat    Prior
##     1.001    normal(0,32)
##     1.000    normal(0,32)
##     1.001    normal(0,32)
##     1.000    normal(0,32)
##     1.001    normal(0,32)
##     1.000    normal(0,32)
##        NA
##        NA
##        NA
##
## Variances:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .prop_vh_15        0.034    0.018        0    0.066    0.034    0.079
##    .prop_vh.f.w_15    5.514    0.950     3.96    7.647    5.514    0.933
##    .prop_vh_17        0.005    0.005        0    0.019    0.005    0.012
##    .prop_vh.f.w_17    6.786    1.161    4.856    9.371    6.786    0.935
##    .prop_vh_19        0.025    0.017        0    0.057    0.025    0.052
##    .prop_vh.f.w_19    6.542    1.115    4.714    9.029    6.542    0.934
##     t1                0.393    0.077    0.265    0.566    1.000    1.000
##    .t2                0.020    0.019        0    0.063    0.043    0.043
##    .t3                0.020    0.017        0    0.054    0.044    0.044
##      Rhat    Prior
```

```
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
```

Property personal

```
## blavaan (0.3-15) results of 2000 samples after 8000 adapt/burnin iterations
##
##   Number of observations                            73
##
##   Number of missing patterns                         1
##
##   Statistic                                 MargLogLik         PPP
##   Value                                       -579.389       0.000
##
## Latent Variables:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t1 =~
##     prop_pers_15      1.000                               0.852    0.996
##     prp_prs.f.w_15    1.000                               0.852    0.321
##   t2 =~
##     prop_pers_17      1.000                               0.896    0.999
##     prp_prs.f.w_17    1.000                               0.896    0.330
##   t3 =~
##     prop_pers_19      1.000                               0.780    0.996
##     prp_prs.f.w_19    1.000                               0.780    0.268
##      Rhat    Prior
##
##        NA
##        NA
##
##        NA
##        NA
##
##        NA
##        NA
##
## Regressions:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t2 ~
##     t1                1.048    0.018    1.013    1.083    0.996    0.996
##   t3 ~
##     t2                0.867    0.015    0.838    0.896    0.996    0.996
##      Rhat    Prior
##
##     1.000    normal(0,10)
##
```

```
##     1.000    normal(0,10)
##
## Intercepts:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .prop_pers_15      3.771    0.100    3.573    3.962    3.771    4.407
##    .prp_prs.f.w_15    6.154    0.310    5.543    6.751    6.154    2.315
##    .prop_pers_17      3.904    0.105    3.695    4.106    3.904    4.351
##    .prp_prs.f.w_17    6.842    0.316    6.208    7.459    6.842    2.523
##    .prop_pers_19      3.812    0.091    3.629    3.989    3.812    4.864
##    .prp_prs.f.w_19    7.072    0.342    6.386     7.74    7.072    2.427
##     t1                0.000                               0.000    0.000
##    .t2                0.000                               0.000    0.000
##    .t3                0.000                               0.000    0.000
##      Rhat    Prior
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##        NA
##        NA
##        NA
##
## Variances:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .prop_pers_15      0.006    0.004        0    0.014    0.006    0.008
##    .prp_prs.f.w_15    6.341    1.076    4.565    8.757    6.341    0.897
##    .prop_pers_17      0.002    0.002        0    0.006    0.002    0.002
##    .prp_prs.f.w_17    6.551    1.115    4.717     9.08    6.551    0.891
##    .prop_pers_19      0.006    0.004        0    0.013    0.006    0.009
##    .prp_prs.f.w_19    7.884    1.362     5.65   10.977    7.884    0.928
##     t1                0.726    0.125    0.522    1.009    1.000    1.000
##    .t2                0.006    0.005        0    0.015    0.007    0.007
##    .t3                0.005    0.004        0    0.013    0.008    0.008
##      Rhat    Prior
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
```

Violence

```
## blavaan (0.3-15) results of 2000 samples after 8000 adapt/burnin iterations
##
##   Number of observations                           73
##
##   Number of missing patterns                        1
```

```
##
##   Statistic                                 MargLogLik         PPP
##   Value                                      -570.826       0.000
##
## Latent Variables:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t1 =~
##     viol_15           1.000                               0.551    0.994
##     viol.f.w_15       1.000                               0.551    0.195
##   t2 =~
##     viol_17           1.000                               0.600    0.990
##     viol.f.w_17       1.000                               0.600    0.194
##   t3 =~
##     viol_19           1.000                               0.553    0.995
##     viol.f.w_19       1.000                               0.553    0.184
##      Rhat    Prior
##
##        NA
##        NA
##
##        NA
##        NA
##
##        NA
##        NA
##
## Regressions:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##   t2 ~
##     t1                1.083    0.028    1.028    1.138    0.994    0.994
##   t3 ~
##     t2                0.916    0.023    0.871    0.962    0.994    0.994
##      Rhat    Prior
##
##     1.000    normal(0,10)
##
##     1.000    normal(0,10)
##
## Intercepts:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .viol_15           2.477    0.065    2.349    2.606    2.477    4.470
##    .viol.f.w_15       5.659    0.333    4.996    6.303    5.659    2.004
##    .viol_17           2.499    0.071    2.357    2.638    2.499    4.124
##    .viol.f.w_17       5.932    0.365    5.217    6.659    5.932    1.922
##    .viol_19           2.533    0.065    2.402    2.661    2.533    4.556
##    .viol.f.w_19       6.357    0.350    5.656    7.039    6.357    2.118
##     t1                0.000                               0.000    0.000
##    .t2                0.000                               0.000    0.000
##    .t3                0.000                               0.000    0.000
##      Rhat    Prior
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##     1.000    normal(0,32)
```

```
##     1.000    normal(0,32)
##     1.000    normal(0,32)
##        NA
##        NA
##        NA
##
## Variances:
##                    Estimate  Post.SD pi.lower pi.upper   Std.lv  Std.all
##    .viol_15           0.004    0.003        0    0.009    0.004    0.011
##    .viol.f.w_15       7.672    1.283    5.567   10.577    7.672    0.962
##    .viol_17           0.007    0.002    0.003    0.012    0.007    0.020
##    .viol.f.w_17       9.168    1.565    6.626   12.732    9.168    0.962
##    .viol_19           0.003    0.003        0    0.009    0.003    0.010
##    .viol.f.w_19       8.703    1.472    6.262   11.958    8.703    0.966
##     t1                0.304    0.053    0.217    0.422    1.000    1.000
##    .t2                0.004    0.003        0    0.011    0.012    0.012
##    .t3                0.004    0.003        0    0.009    0.012    0.012
##      Rhat    Prior
##     1.001 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.000 gamma(1,.5)[sd]
##     1.001 gamma(1,.5)[sd]
##     1.001 gamma(1,.5)[sd]
```

Include stable method effect

Additionally, we can expand this model and include a stable method effect. Now the standardized loading on the method effect is "systematic bias" while the standardized loading on trait can be seen as "reliability" I think. These models are getting hard to estimate so we need to trade lightly.