Milojev, P., Osborne, D., & Sibley, C. G. (2014). Default statistical standards and modelling decisions for NZAVS data analysis. *NZAVS Technical Documents, e11.*

This document is an internal 'living' document outlining our statistical and analytic defaults. The NZAVS is a complex longitudinal and extremely large scale dataset. As such, we generally follow recommended best practise in our area, but find ourselves reporting more and more models for which general standards in our area are unavailable. As such, this document provides an explicit statement about our internal team decisions and recommendations for analysis of NZAVS data. This is useful for reviewers and readers who are interested in our defaults. It is also important for maintaining a general standard and level of integrity in analyses across our broad and diverse research team.

Document History:

16 April 2014 – Last updated

14-April-2014 – Revised following discussion with broader NZAVS team

12-April-2014 – Document created

1) Syntax for all models is posted on the NZAVS website, and data used in our models is available upon request for the purposes of replication or further joint collaboration. We run the majority of our analyses in Mplus. This is our preferred analysis package.

2) When determining statistical significance, practical significance, and data screening…

    i) In models using frequentist estimates (e.g., ML), we generally take $p < .01$ as our default for statistical significance (implemented 12-April-2014). For NZAVS analyses using a frequentist approach, we view $p < .05$ as marginally significant, and below our official standard (again, implemented 12-April-2014).
    ii) When estimating frequentist models, we use two-tailed p-values by default, but see point 5 below about our general default being to use Bayesian estimation.
    iii) We report effect size estimates whenever possible (standardized betas in most cases, variance explained, ORs in logistic regression models, etc).
    iv) We routinely check for variance inflation using the VIF, and potential violation of assumptions about a given distribution, and heterogeneity of residuals, etc (see also 7.i).

3) When dealing with missing data...

i) When income is included as an exogenous variable in the model (and hence listwise deletion would have applied by default) we replace missing values with the median (but have previous to 12-April-2014 been replacing with the mean). This is always stated, and missing values for other exogenous manifest variables in regression models are not replaced by default.
ii) In most models, and unless specifically interested in income, we will preference including the NZDep index rather than household income. This is because the NZDep has more complete information (i.e., less missing data). Our default is to include the NZDep at the participant level, unless the focus of the paper is on other multilevel parameters (given that we tend to only have 1.x people per meshblock area unit.
iii) In latent variable models, missing data for manifest indicators is estimated using FIML (the default Mplus setting).
iv) In latent growth models, our default is to limit models to include participants who have completed a minimum of two waves. Missing data in LGMs is estimated using default Mplus settings. Again, this will always be stated in reported analyses.
v) If other methods of missing data imputation are used (and we expect these will become more common with time), this will always be stated.

4) We have a default list of standard demographic control variables or 'usual suspects' that we typically adjust for, unless we have a specific reason not to. This is fairly flexible, and our default covariate list is more likely to be included in less computationally intensive models (such as regression). It also depends on theoretical relevance and focus (and in some cases beyond our control the journal word limit) for a given model/paper. As a very general default, we tend to run models controlling for…

   i) Gender, Age, NZDep, Religion, Relationship status, Parental Status, Employment, Education, Ethnicity (often majority European versus all other), Urban-Rural (this latter introduced as a default on 12-April-2014).
   ii) In models employing cross-sectional analyses of latter time points, a code identifying original sample frame versus booster sample should also be included as a default control.
   iii) These controls may not always be included in reported correlation matrices (due to space constraints), but will always be provided upon request. However, we do aim to always include full correlation matrices whenever feasible.
   iv) Generally, models with focal predictors will be included as the first step, and then the model will be replicated showing that it holds with default covariates, and any other specific covariates relevant to that model. This is most often the case for regression-type models, and not more complex models where processing or data constraints make it difficult to include a large number of 'usual suspect' covariates (such as in a latent growth model, or models with many random effects). An

explicit statement will always be included stating which, if any, covariates have been included.

v) When including income or other variables with a very large range, we explicitly state if a transformation or scaling factor has been applied (our standard scaling factor for income is /10,000 for linear rescaling, or to take LG10 of income).

5) We use Bayesian analysis as our primary default method of estimation unless we have a specific reason not to. Reasons not to employ a Bayesian approach generally include…

   i) In cases where standard fit indices are of interest, such as CFAs;
   ii) In cases where we are asked to change to ML models during the review process,
   iii) In cases where a Bayesian approach is not applicable or not implemented in Mplus, such as some multilevel models, latent variable interaction models, some mixture models.

6) When a Bayesian approach is used…

   i) Given the large sample size of the NZAVS (and hence most of our analyses), our preference is to use non-informative priors for fixed effects with Mplus default settings (e.g. $b \sim N(0, 10^{10})$). We will always explicitly state if informative priors are used, and state what those priors are. (Note. when you condition on essentially infinite variance components you recover most of the properties of a frequentist estimator. However, we prefer a Bayesian estimator because of the way in which it pools information across the entire model).
   ii) When assessing fit, to always report the BIC and posterior predictive p-value, and chi-square and 95% CI for the posterior predictive p-value.
   iii) When informative priors are used, our default is to use weakly informative priors where the variance of the prior is multiplied by a factor of 4 and assumed to be normally distributed, unless there is a specific reason otherwise. This generally will have the effect of treating the prior estimate as if it were based on a quarter of the sample size from which it was derived.
   iv) To always report unstandardized and standardized betas (where possible) posterior SDs, and the 95% credible intervals around b. Point estimates for the posterior distribution are the median by default.
   v) In addition to credible intervals, we also report one-tailed pMCMC p-values for parameters in Bayesian models. These represent the proportion of the posterior distribution for an effect that is below 0 (if the effect is positive) or above 0 (if the effect is negative).
   vi) We consider the Region Of Practical Equivalence (ROPE) for most questions relating to the NZAVS to be a credible interval or highest density interval where 95% of the posterior distribution is within -.05 to .05 of the point estimate.

vii) When estimating CFA or general SEMs under Bayes, by default we model the residual covariance of all manifest indicators with a normal distribution, around 0 with variance of $\sigma^2 = .01$. This is always stated when done.
viii) To always estimate our models with visual inspection of the posterior distributions and chains as a quality control, and then doubling all FBITERATIONS, chains, etc, beyond those needed to pass the Mplus PSR defaults and our own visual inspection. Generally, we set chains = 8 and FBITERATIONS = 50,000.
ix) We strongly prefer to estimate indirect effects, where possible, using a Bayesian approach, and report 95% values for the HDI of the indirect effect, rather and a frequentist bootstrapping approach.

7) When not using a Bayesian approach for standard regression and SEM models…

i) Our default for regression models will be to use MLR, and when MLR is used to always check that the model holds when using ML, and note any discrepancies that might alter conclusions. When using a frequentist analysis, we prefer MLR because it is more robust for minor violations of assumptions regarding distribution of residuals.
ii) When estimating model fit in multi-group CFA, our default is MLR.
iii) When estimating model fit in standard SEMs, we employ ML as standard fit indices are not available in Mplus under MLR (outside of multi-group models).
iv) To report unstandardized and standardized betas (where possible), SE(b) two-tailed p, and the 95% confidence interval around b. Likewise, to report b, se, OR and CI(b) for logistic models.
v) When assessing indirect or mediation effects in frequentist models, to report bias corrected bootstrap CIs with 5000 bootstrap resamples.

8) When assessing model fit under ML or MLR (multi-group)…

i) We will always report the chi-square, df, sRMR, RMSEA, and CFI, and 95% CI for the RMSEA, BIC and AIC. (AIC implemented 12-April-2014).
ii) We generally take the following values as indicating reasonable model fit. sRMR < .08, RMSEA < 06, CFI > .95.
iii) We report all fit indices to 3 d.p. by default.
iv) To always inspect the modification indices, and explicitly state if any additional post-hoc parameters have been included on the basis of these.

9) When estimating standard (typically cross-sectional) Structural Equation Models …

i) See point 8.i. about assessing model fit.

ii) When including covariates, to model them as exogenous variables at the same level as the most distal focal predictors with direct effects on all endogenous variables.
iii) To explicitly state if any additional parameters have been included (such as correlated residuals). (An obvious point in our view, but concerning when this is sometimes not done in the literature).
iv) Statement about modification indices see 8.iv

10) When estimating Multilevel Models or Hierarchical Linear Models with participants nested within regions…

i) To use either census area units (n of ~1400) or territorial authorities (n of ~70) as the default areas within which participants are nested. CAUs provide a multilevel structure where variation across small geographical units is the theoretical focus, territorial authorities for when differences across broader regions are of interest.
ii) Note that (i) above does not apply when political attitude and voting behaviour are of interest. In that case, we will use registered electoral districts as the default higher level unit of interest.
iii) To explicitly state which centring options are used (grandmean versus group mean). For most of our focal questions of interest grandmean tends to be our preferred option, as our question about people-within-regions tend to focus on deviation from the national average, rather than deviations from the average within regions. (We are aware that there is disagreement about this, and that for many focal research questions, groupmean centring tends to be preferred).
iv) To include random effects for all focal predictors where possible (this generally depends on model complexity, but we assume random effects by default).
v) To include random effects for covariates when this is also possible, but in complex models to allow fixed effects estimates for covariates that are not of focal interest.
vi) To report between and within variances and random effects for all parameters (unless asked to remove these to save on space, etc). (implemented 12-April-2014).

11) When estimating Mixture Models (LPA, LCA, mixture regression, etc)…

i) When assessing model fit, to always report the BIC and AIC for a range of fit indices from 1 to (preferred + 2).
ii) When estimating model fit, to also be based on visual inspection of the solution and take care not to simply extract classes or profiles are that incrementally higher or lower by a similar amount across indicators.
iii) When estimating LPA with covariates, we avoid the single-step estimation of covariates, and employ the three-step or distal three-step procedure so that covariates do not influence the LPA itself.
iv) When the focus is on small classes (less than 5% of the sample) to take particular care to check the stability of these small classes by

inspecting over-extracted classes that should start to run into point (ii).

v) To double the STITERATIONS once a stable model has been achieved.

12) When assessing moderation (regression interactions)…

i) To include SEs, or CIs around the point estimates on graphs.

ii) To solve the simple slopes at practically meaningful values wherever possible (this is consistent with recommendation from Aiken and West, 1991). To solve for slopes at +/-1 SD of the moderator only when more practically meaningful values are not feasible, e.g., likert scale scores).

iii) To generally report tests of differences between points at low and high values of the predictor, as well as the significance of slopes at conditional values of the moderator.

iv) To always explicitly state our reasonable range of values of the predictor when solving curvilinear interaction effects, and base these explicitly on the distribution of values of the predictor (e.g., not going beyond reasonable value in the data for curvilinear effects of variables such age).

v) When the focal research question is about assessing curvilinear effects, to use a 'work-backwards' approach. This means, for example, including the higher-order factorials at the first step, and then removing the cubic if non-significant, then the quadratic if also non-significant, before claiming a linear interaction. In contrast, when the focal research question is about assessing a linear interaction, to stick with a linear interaction (although in this latter case, we will also check for curvilinear effects even if not reported).

13) When estimating Latent Growth Curve models…

i) We will always report a baseline model checking for significant variance in the slope and intercept before adding covariates.

ii) To generally employ time-varying random effects where possible (i.e., allowing for differences across people in the lag between waves). Note that this is not always feasible.

iii) When using equidistant assessments, our default will be to fix at values at yearly intervals, i.e., where a change of one-unit represents one year.

iv) To always report the overall slope and intercept, residual variances in the intercept and slopes, and the correlation between the intercept and slope.

v) To generally present a 'three-graph' package as standard, where we have a general graph presenting the trajectories at conditional values of the covariate of interest, along with graphs presenting the intercept and slopes at a the full number/range of conditional values. And to also include HDIs or SEs for all estimates in these graphs.

vi) By default, to include mean-centred covariates, so the overall slope represents the rate of change at the mean. To always explicitly state the centring option that has been employed.

14) When estimating cross-lagged models…

i) Where possible, and when we have more than two or three time points, latent growth models, latent difference score growth models, or dynamical models of change, etc, are to be preferred. However, in some cases, we think that standard cross-lag models can be useful.
ii) Variables will be modelled as latent using all available indicators, i.e. not at mean scores. This adjusts for measurement error across waves.
iii) We will generally include models testing the reverse cross-lagged paths, even if only in a footnote or appendix.
iv) Covariances will be allowed among residual variances of latent variables within time. In our view, when testing cross-lag models using panel data, allowing the correlation of residual variances generally provides a more robust test of the cross-lagged effects.
v) In full cross-lag models (i.e., all variables at both times), models constraining the stability of variables across time to equality will also be tested.
vi) A test of directionality would involve showing that one cross-lag path is larger than the reverse, when constraining the stability of both latent variables over time to be equal and allowing within-time covariance of residuals.

15) When estimating Markov Chain Models or Latent Transition Models…

i) To generally assume a non-stationary model for the first wave. This recognizes that the base-rate may differ from the rate of change at t + 1, especially when including covariates to predict transition probabilities.
ii) To generally model Markov chains with observed variables at latent or 'hidden' Markov models, so as to adjust for measurement error.
iii) To always report all transition probabilities, including those for all covariates.
iv) To report BIC, chi-square, and likelihood ratios for all models.