A Mega-Analysis of Personality Predictions: Robustness and Boundary Conditions

Emorie D Beck[1]

[1] Washington University in St. Louis

23. January 2020

Study Information

Title	A Mega-Analysis of Personality Predictions: Robustness and Boundary Conditions

Description	Personality traits are relatively stable, dispositional patterns that differentiate people from one another (Roberts, Wood, & Caspi, 2008). Moreover, personality traits predict many important life outcomes, such as marriage (Kelly & Conley, 1987; Malouff, Thorsteinsson, Schutte, Bhullar, & Rooke, 2010; Specht, Egloff, & Schmukle, 2011), life expectancy (Jackson, Connolly, Garrison, Leveille, & Connolly, 2015; Jokela et al., 2013; Martin, Friedman, & Schwartz, 2007; Turiano, Chapman, Gruenewald, & Mroczek, 2015), and health (Hampson, 2012; Weston, Hill, & Jackson, 2015). Much research has examined what personality predicts (see Ozer & Benet Martinez, 2006, and Soto, 2019, for reviews), including what the strongest predictors of different outcomes are (see Roberts, Kuncel, Shiner, Caspi, & Goldberg., 2007).
	Despite numerous studies and meta-analyses that demonstrate personality prospectively predicts consequential life outcomes, most studies fail to capture true selection

effects for two main reasons: reverse causality as well as a combination of researcher degrees of freedom, publication bias, and a lack of understanding of boundary conditions. Reverse causality could occur because many of outcomes that personality prospectively predicts also prospectively predict personality. For example, Conscientiousness predicts both work success (Judge, Higgins, Thoresen, & Barrick, 1999; Rothmann & Coetzer, 2003) and health (e.g. Weston, Hill, & Jackson, 2015). However, socioeconomic status also predicts both work success (e.g. Ensminger, Fothergill, Bornstein, & Bradley, 2003) and health (e.g. Adler, Boyce, Chesney, Folkman, & Syme, 1993), and work success and health also predict each other (e.g. Conger & Donnellan, 2007). Thus, the direction of the relationships among personality, consequential outcomes, and potential moderators make it difficult to tease apart their relationships. Thus, the existent work examining selection effects of personality on life events cannot disentangle selection effects of personality from selection bias of other background characteristics, like health, work factors, and socioeconomic status, which may explain why evidence of personality predicting life events has been mixed. In other words, people who experience the life events that personality predicts might differ from those who do not in important ways that also influence personality.

In the present study, using 10 longitudinal panel studies, I will examine whether 14 measures of personality at a baseline assessment predict the future experience of 14 life events and outcomes that occur throughout the lifespan independently of selection bias of other background characteristics and of procedural and analytic procedures. Previous work has examined selection effects of personality on life events using these samples, but the present study (1) tests whether selection effects persist after accounting for background characteristics via propensity score matching and (2) specification curve techniques to examine the different boundary conditions where personality-outcome associations exist, each (3) using a mega-analytic procedure to test the robustness of effects.

Research Questions

1. Do personality traits predict consequential outcomes:

- without controlling for selection bias?

- controlling for selection bias?

- robustly when systematically controlling for theoretically consequential covariates?

Design Plan

Study type	**Observational Study**. Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, natural experiments, and regression discontinuity designs.
Blinding	No blinding is involved in this study.
Study design	All of the studies in the present investigation are large scale, longitudinal panel studies detailed more below.
	The basic structure of the study involves using 14 personality variables (e.g. the Big 5, self-esteem) to predict 14 life events and outcomes (e.g. child birth, retiring), while testing for 6 moderators (e.g. SES, gender) using both propensity score matching to control for selection bias (study 1) and specification curve analysis to test the robustness to covariates (study 2) in 10 longitudinal studies.
Randomization	Some studies randomized which participants received certain questionnaires. Where possible, we are avoiding the use of such data because it restricts sample sizes.

Sampling Plan

Existing data	**Registration prior to accessing the data**. As of the date of submission, the data exist, but have not been accessed by you or your collaborators. Commonly, this includes data that has been collected by another researcher or institution.
	Applicable for the following data sets: WLS MIDUS Add Health LISS

Registration prior to analysis of the data. As of the date of submission, the data exist and you have accessed it, though no analysis has been conducted related to the research plan (including calculation of summary statistics). A common situation for this scenario when a large dataset exists that is used for many different studies over time, or when a data set is randomly split into a sample for exploratory analyses, and the other section of data is reserved for later confirmatory data analysis.

Applicable for the following data sets:

HILDA

SHP

BHPS

Registration following analysis of the data. As of the date of submission, you have accessed and analyzed some of the data relevant to the research plan. This includes preliminary analysis of variables, calculation of descriptive statistics, and observation of data distributions. Please see cos.io/prereg for more information.

Applicable for the following data sets:

HRS

GSOEP

CNLSY

Explanation of existing data

I have never previously worked with the WLS, MIDUS, Add Health, or LISS data. I have only accessed their codebooks online.

I have previously worked with life satisfaction data from BHPS and SHP and worked with an undergraduate studying the relationship between satisfaction and goals in the HILDA study.

I have previous examined selection and socialization of the Big 5 and health events (HRS) and life events (GSOEP) and other individual difference variables and criminal behavior in adolescence (CNLSY).

I have never worked with any of these in a mega-analytic framework.

Data collection procedures and Sample Sizes

Ad Health

The National Study of Adolescent to Adult Health (Ad Health; Harris & Udry, 2018) is an ongoing longitudinal study of adolescents in the United States that began as a response to a federal mandate to better understand adolescent health. The data are available online at https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/21600.

The initial sample of participants included approximately 20,000 students who completed at home administrations the study. Four waves of data collection (1994-1995, 1996, 2001-2002, and 2008) have been completed. The latest release contains data through 2008. Another wave of collection began in 2016 but has not yet been released. More documentation of the data are available at https://www.icpsr.umich.edu/icpsrweb/content/DSDR/add-health-data-guide.html#intro.

Sample sizes vary by year, from 14,738 (1996) to 20,745 (1994-1995). This provides 99% power to detect a correlation effect size of ~.03.

###BHPS

The British Household Panel Study (BHPS; University of Essex, 2018) is a longitudinal study of households in the United Kingdom. These data are available online, through application, from https://www.iser.essex.ac.uk/bhps/about/latest-release-of-bhps-data.

Participants were recruited from more than 15,000 individuals from approximately 8,000 households in the United Kingdom. Data have been collected annually since 1991 from approximately 10,000 individuals (5,500 households) in Great Britain but expanded to include Scotland and Wales in 1999 and Northern Ireland in 2001. In 2010, the BHPS stopped data collection, but 6,700 of the current 8,000 participants were solicited to become part of the broader Understanding Society study (University of Essex, 2019). Participants can be matched across studies, so I will use additional data on the original BHPS participants from the Understanding Society study for additional waves of outcome data.

Sample sizes vary by year, ranging from 10,264 (1991) to 14419 (2008). This provides 99% power to detect a zero-order correlation effect size of ~.05, two-tailed at alpha

.05.

GSOEP

The German Socioeconomic Panel Study (GSOEP; Socio-Economic Panel, 2017) is an ongoing longitudinal study of German collected by the German Institute of Economic Research (DIW Berlin). The data are freely available at https://www.diw.de/soep by application.

Data have been collected annually since 1984 (the latest data release includes data up to 2017). Participants have been recruited from more than 11,000 households, which are nationally representative of private German households. 20,000 individuals are sampled each year, on average. It is critical to note that the GSOEP samples households, not individuals, and the households consist of individuals living in both the "old" and "new" federal states (the former West and East Germany), foreigners, and recent immigrants to Germany.

Sample size varies by year, ranging from approximately 10,000 (1989) to 31,000 (2013). This provides 99% power to detect a zero-order correlation effect size of ~.06, two-tailed at alpha < .05.

HILDA

The Household Income and Labour Dynamics in Australia (HILDA; Wilkins, Laß, Butterworth, & Vera-Toscano, 2019) study is an ongoing longitudinal study of Australian households. These data are available through application from https://melbourneinstitute.unimelb.edu.au/hilda/for-data-users.

Participants were recruited from more than 17,000 individuals. Data have been collected annually since 2001. The latest data release includes 17 waves of data from 2001 to 2017. More documentation can be found in the HILDA data dictionary at https://www.online.fbe.unimelb.edu.au/HILDAodd/srchSubjectAreas.aspx.

Sample sizes vary by year, ranging from 12,408 (2004) to 17,693 (2016). This provides 99% power to detect a zero-order correlation effect size of ~.03, two tailed at alpha .05.

HRS

The Health and Retirement Study (HRS; Juster & Suzman, 1995) is an ongoing longitudinal study of households in the United States. These data are available at https://hrs.isr.umich.edu by creating a free account.

Participants were recruited from more than 35,000 individuals from the financial households of individuals born between 1931 and 1941 in the US. Data have been collected biannually since 1992. The latest data release includes data up to 2016. On average, 10,000 individuals are sampled each wave More information on the HRS can be found at https://hrs.isr.umich.edu/documentation/survey-design, but, in short, the HRS is a nationally representative sample of adults over 50 in the US. It is critical to note that the HRS samples households of the original cohort and follows individuals and their spouses or partners until their death.

Sample size varies by year, ranging from approximately 7,500 (2014) to 15,500 (1992). (https://hrs.isr.umich.edu/sites/default/files/biblio/ResponseRates_2017. pdf). This provides 99% power to detect a zero-order correlation effect size of ∼.04, two-tailed at alpha .05.

LISS

The Longitudinal Studies for the Social sciences (LISS; Scherpenzeel, Das, Ester, & Kaczmirek, 2010) is an ongoing longitudinal study of households in the Netherlands. These data are online, through application, from https://statements.centerdata.nl/ liss-panel-data-statement.

Participants were approximately 8,000 Dutch-speaking individuals permanently residing in the Netherlands from 5,000 households. Data have been collected annually since 2007. The latest data release includes 11 waves of data from 2008 to 2018. More documentation are available at https://www.dataarchive.lissdata.nl/ study_units/view/1.

Sample sizes vary by year, ranging from 5,021 (2018) to 6808 (2008). This provides 99% power to detect a correlation effect size of ∼.04, two-tailed at alpha .05.

MIDUS

The Midlife in the United States (MIDUS; Brim, Ryff, & Kessler, 2004; Ryff et al., 2012, 2016) study is an ongoing longitudinal study of adults in the United States. These data are available at http://www.icpsr.umich.edu by making a free account.

Participants included more than 10,000 individuals aged 25 or older from the United States. The present study uses data from MIDUS I, II, and III. MIDUS I was collected in 1995-1996. MIDUS II was the follow-up to MIDUS I and was collected from 2004-2006. MIDUS III was an additional follow-up conducted from 2013-2014. More information can be found at http://midus.wisc.edu/findings/Understanding_ Data_Collection_in_MIDUS.pdf.

Sample size varies by wave, with 7,108 (MIDUS I), 4,963 (MIDUS II), 3,294 (MIDUS III). This provides 99% power to detect a zero-order correlation effect size of ~.06, two-tailed at alpha .05.

NLSY

The Children to Young Adults Study (CNLSY; Bureau of Labor Statistics, 2017) is an offshoot study of the National Longitudinal Study of Youth (NLSY79), which is an ongoing longitudinal, nationally representative study of more than 12,500 individuals in the United States that began in 1979. The CNLSY includes the biological children of the NLSY79 participants and began in 1986. Children (10 years and older) completed separate inventories from children (or "young adults") aged 15 and above. Mothers of children 10 and below also completed surveys on the children prior to age 10. All participants were interviewed in addition to surveys.

Sample sizes vary by year, ranging from approximately 1,331 (1979) to 11,530 (2016). This provides 99% power to detect a zero-order correlation effect size of ~.05.

The present study included data from a subsample of 7,736 CNLSY participants who had at least one wave of measurement for each of the following: matching variables (see Supplementary Materials; partial data allowed), life event / outcome variables (e.g. marriage), and personality variables (e.g. the Big Five).

SHP

The Swiss Household Panel Study (SHP; Voorpostel et al., 2016) "Living in Switzerland" is an ongoing longitudinal study of households in Switzerland. These data are available online, through application from https://forsbase.unil.ch/project/study-public-overview/15632/0/.

Participants were recruited from more than 10,000 individuals from the households whose members represent the non-institutional resident population of Switzerland. Data have been collected annually since 1999. The latest data release includes data up to 2018. On average, about 5,000 individuals are sampled at each wave. More documentation can be found at their website, but, in short, the SHP is a nationally representative sample of Swiss citizens.

Sample sizes vary by year, ranging from 5,220 (2003) to 13,295 (2013). This provides 99% power to detect a zero-order correlation effect size of ∼.06, two tailed at alpha .05.

WLS

The Wisconsin Longitudinal Study (WLS) is an ongoing longitudinal study of individuals who graduated from Wisconsin high schools in 1957 and were born between 1937 and 1940 as well as their siblings.

Graduates were randomly recruited from Wisconsin high schools in 1957 and born between 1937 and 1940. In 1977, at least one sibling of the original graduates from 2,100 families were also invited to participate in the study. As such, the study is representative of older, white Americans who have at least a high school education. Graduate data have been collected in in 1957, 1964, 1975, 1992, 2004, and 2011, and sibling data have been collected in 1977, 1994, 2005, and 2011. Personality data were initially collected in 1992 for graduates and 1994 for siblings. More documentation can be found at https://www.ssc.wisc.edu/wlsresearch/.

Sample sizes vary by wave, from 9,681 (2011) to 10,317 (1957). This provides 99% power to detect zero-order correlation effect sizes of ∼.06, two-tailed at alpha .05.

Sample size rationale	Sample sizes are restrained by (1) available data and (2) the timing of the personality measures.

Variables

All variables, their documentation, recoding (in `R` code) and composite rules have been recorded in the codebook attached with this preregistration, but can be broken down into 5 categories:

1. Procedural Variables: Participant and Household ID's as well as ID's of parents for some studies.

2. Matching Variables: matching variables for propensity score matching.

3. Personality Variables: 14 variables representing personality indicators

4. Outcome Variables: 14 outcome variables representing major life events and outcomes

5. Covariates: key theoretical variables for specification curve analysis.

Measured variables	In the proposed study, I will test how 14 personality characteristics are associated with 14 life events and outcomes, while controlling for more than 50 background (matching) characteristics (propensity score matching) or covariates and operationalizations (specification curve analysis), testing seven potential moderators of personality predicting such outcomes. For a full overview of which personality, outcome, and moderator measures are available across data sets, see Table 1 in the attached "tables.xlsx" file.

Personality

I will examine 14 personality characteristics: the Big Five (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience), self-esteem, optimism / pessimism, subjective well-being, locus of control, social support, life

satisfaction, negative affect, positive affect, depression, and intelligence. Full information on the scales used for each of these measures for each study is presented Table 2 in the attached "tables.xlsx" file. Because none these data sets have been accessed in the design of this study, descriptive statistics for each scale are not available at this stage. However, as is clear in Table 2, many of the measures were on different scales, so all personality indicators will be operationalized as Percentages Of the Maximum Possible score (POMP) in the mega-analytic procedure (Cohen, Cohen, Aiken, & West, 1999). Unlike standardization procedures, that have a mean of zero and unit variance and can be misleading when data are skewed, POMP does not rescale sample variance based on the observed data, which overly relies on deviations from the mean. Instead, POMP relies on the ratio between the difference between a score and the minimum and the maximum and minimum, or

$POMP = \frac{observed - minimum}{maximum - minimum}$.

Life events and outcomes

I will investigate whether personality predicts 14 life events and outcomes. A full list of outcomes can be seen in Table 2. The outcomes chosen are those frequently tested in other studies and most likely to be included in large panel studies. For each outcome, participants reported whether the outcome had occurred in the survey year or years prior or, in some cases, the year an outcome or event first occurred. Responses will be coded as "1" for that event if participants reported experiencing it anytime between the year after the utilized personality measure to the latest available wave and "0" otherwise. Participants who experienced events prior to the first personality measure will be excluded.

Moderators

In addition to matching for common demographic variables, like age and gender, I will additionally test whether these, as well as survey year, ethnicity, socioeconomic status (SES), personality measure reliability, and study moderate the relationship between personality and outcomes. Age and SES will be coded at the time of measured personality. Age, and SES will be test as continuous Level 1 moderators, while ethnicity and gender will be tested as binary Level 1 moderators. Reliability

and survey year will be tested as continuous Level 2 moderators. The effect of study will be tested by examining the Level 2 Variance of personality-outcome associations. Significant variance will indicate that study moderates the effect.

Matching variables and specification curve covariates

Matching variables are those to be used in the propensity score analysis to match those who did or did not experience different life events or outcomes. Target variables can be roughly broken down into eight categories: demographics (e.g. sex), activities (e.g. volunteering), financial (e.g. gross wages), household (e.g. number of household members), health (e.g. BMI), psychological (e.g. life satisfaction), relationship (e.g. relationship with father), and social (e.g. visits to friends). In order to construct more reliable measures and not exclude participants who did not respond to surveys in the same year as the personality measures, matching variables were pooled across time, using all available data from the earliest wave of the study to the year of the utilized personality measure for each person. Cross-study covariates will be operationalized using POMP, with the exception of core variables with meaningful scales (e.g. age, gender, etc.). The full details of the construction of these variables are available in the codebook.xlsx file attached with this preregistration and a breakdown of covariate availability across studies is available in Table 3 in the tables.xlsx file.

Analysis Plan

Statistical models

Study 1: Propensity Score Matched mMega-analysis of longitudinal studies.

Confirmatory analyses will be tested using a series of multilevel Bayesian logistic regression models implemented using the `brms` (Bürkner, 2017, 2018) package in `R` (R core team, 2018). I will use the default priors for all fixed effects, which are "uninformative" priors meant to regularize the models. However, given the sample sizes in the present model, the data are likely to overwhelm these priors. I will use

half Cauchy priors for Level 2 variances and LKJ Cholesky covariance priors for Level 2 covariances.

The analysis phase will consist of three main parts, with interim steps to link these together: multiple imputation, propensity score matching, and tests of selection effects using multilevel Bayesian logistic regression models.

First, I will use multiple imputation to impute missing data for the matching variables separately for each study. Before doing so, I will first create composites of the matching variables prior to the year for which we will use personality data for each study. I elected to use composites rather than survey responses from the year prior to personality measures due to irregularities in survey construction and responses that would severely restrict the number of observations. To ensure transparency, I will conduct all analyses using the raw data imported directly from the data files obtained from data maintainers for each study, and all steps in creating the composites are documented in an extensive codebook containing the item lists, text, scales, and recoding of all variables for all studies. Moreover, all steps will be documented in files and code shared on the Open Science Framework and GitHub.

The composite matching variables will then be used in multiple imputation and propensity score matching, which requires completely non-missing data. Multiple imputation will be conducted using the `amelia` package in `R` (Honaker, King, & Blackwell, 2011). I will impute 10 data sets.

Second, I will use the multiply imputed data to calculate propensity scores for each of the multiply imputed data sets for each outcome and study separately. In addition, I will conduct the procedure separately for each individual difference characteristic, such that each propensity score matched set will be matched on all matching variables and all personality variables except for the target variable that will be used to predict each outcome, making this an extremely conservative test of the relationship between personality and outcomes. Finally, separate propensity score matched sets will also be generated to test each of the Level 1 moderating questions (age, gender, SES, and ethnicity).

The propensity score matching procedure attempts to equate those who did or did not experience an outcome by assigning each person a risk score based on a number of background factors. Then each person who experienced the outcome is matched with someone else in the control group who had a similar "risk" of experiencing the

outcome. Matching will be done using the matchit packages in R (Ho, Imai, King, & Stuart, 2011). Because the sample sizes of the groups of people who experience specific outcomes are much smaller than the individuals who did not experience them, we choose to use propensity score matching rather than propensity score weighting. I will begin by using "nearest neighbor" matching and a ratio of 2 to 1 and a caliper width of .25 σ (Guo & Fraser, 2015) and iteratively increase the ratio for outcomes that were not balanced using these criteria.

Third, I will test for selection effects using a series of multilevel Bayesian logistic regression models using the brms package in R (Bürkner, 2017, 2018). In all models, the "no outcome" group will be considered the reference group. Using unmatched data sets and matched data sets that did not account for personality or moderators at the wave of personality assessment, I will predict outcomes from baseline personality. With the exception of moderator analyses, I will not include additional covariates (e.g. age, gender) in these models because these should be effectively controlled for in the propensity score matching procedure. The basic form of the model is as follows:

Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} * P_{ij} + \varepsilon_{ij}$ Level 2:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j},$$

where γ_{00} is the average log odds of experiencing the outcome across all studies and γ_{10} multiple of log odds change associated with a one-unit change in the percentage of the maximum of the possible (POMP) personality score. All results will be presented both as log odds and as odds ratios (OR) with \89% uncertainty intervals (UI). u_{0j} indicates the difference between the average estimate of log odds of experiencing an outcome and the estimate for each study (i.e. the study-specific estimate of the log odds of each outcome), and u_{1j} indicates the difference between the average multiple of log odds associated with a one unit change in POMP personality score and the estimate for each study (i.e. the study-specific estimate of the personality-outcome relationship). Each of these will be presented as forest plots showing both study-specific and average effects. Moderator analyses will extend the form of the core analyses by additional terms at Level 1 ($\beta_{2j} * moderator_{ij}$ and $\beta_{3j} * P_{ij} * moderator_{ij}$; age, ethnicity, gender) or Level 2 ($\gamma_{01} * moderator_j$ and $\gamma_{11} * moderator_j$; reliability, survey year, study). For Level 1 moderators, I will include random effects that will capture the study-specific effects for each moderator term.

Study 2: Specification curve analysis of longitudinal studies

As a second test of the robustness of prospective personality-outcome associations, I will conduct a specification curve analysis for each personality-outcome combination (Simonsohn et al., 2015).

Specification curve analysis is carried out in three main steps. First, the researcher defines the set of reasonable model and variable specifications. Second, the researcher estimates all of these reasonable specifications and represents them using a specification curve. Finally, the researcher constructs an inferential specification curve using join statistical analyses (Simonsohn et al., 2015). Constructing the specification curve in the second step serves both to show the full range of how specifications influence the statistical results as well as which specifications are most consequential for the results, while using the specification curve inferentially in the third step allows the researcher to make a statistical inference about whether the curve is inconsistent with a null hypothesis of no effect of personality on outcomes. This is most simply done with a permutation test in which the consequential variable (in the context of this paper, whether an outcome occurred for a participant) is shuffled. Shuffling the consequential outcome and re-estimating the full set of possible specifications, except now when there is no reason to expect a relationship between predictors or covariates and outcomes, produces a distribution of specification curves under the null of no personality-outcome relationship.

Defining specifications

The first step in specification curve analysis involves defining the set of valid specifications. Because I am testing the relationship between different personality characteristics and outcomes, the set of valid specifications will differ across outcomes. Thus, a unique set of specifications was established for each outcome separately based on (1) covariates used in previous studies to predict those outcomes, (2) other covariates that I identified as plausibly related to each outcome based on theory, (3) using the other personality characteristics that are not the focal part of the target personality-outcome association, and (4) different operationalizations of covariates used in previous studies. Different operationalizations include whether variables are treated as continuous or binarized or trichotomized, as was sometimes done in previous studies. In addition, the chosen specifications include whether

the covariates are cross-sectional (i.e. measured in the same year as personality) or composites of multiple waves of measures of the covariate (i.e. measured prior to personality). Because the focus is on covariates that may predict the outcome, the same set of covariates will be used to test all personality-outcome associations for each outcome. Table 3 in the attached "tables.xlsx" file presents the full set of covariates and operationalizations that will be used for each outcome, as well as the total number of specifications that result from the set of specifications.

Defining the specification curve

The next step in specification curve analysis is to run the model, specifying all combinations of the specifications from the previous step. Once this is done, the target personality-outcome association is extracted from each model and ordered from strongest negative to strongest positive to define the specification curve. Thus, I will define a total of 196 specification curves (14 personality characteristics x 14 outcomes). Although I could compute separate models for each study, the number of specifications would be so large as to be intractable, so I will restrict the model to multilevel logistic regression. Because conducting the models in a Bayesian framework would require resources likely well beyond the scope of this project, I will fit in a frequentist multilevel model framework using `lme4`.

The basic form of the model is the same as for the propensity score matched mega-analysis, but with additional covariates (β_{2j} to β_{pj}): Level 1: $Y_{ij} = \beta_{0j} + \beta_{1j} * P_{ij} + \beta_{2j} * X_{1ij} + \ldots + \beta_{pj} * X_{pij} + \epsilon_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + u_{0j}$ $\beta_{1j} = \gamma_{10} + u_{0j}$ $\beta_{2j} = \gamma_{20}$ $\ldots$ $\beta_{pj} = \gamma_{p0}$, where γ_{2j} to γ_{pj} represent Level 1 main effects with no random slope (i.e. study level slope).

Permutation-Based Inferential Test

The final step in specification curve analysis involves conducting a permutation-based test to determine whether the observed specification curve differs from the null of no relationship between any predictors and the outcome. Thus, the outcome variable is shuffled and the specification curve procedure from the second step is repeated a large number of times. The observed specification curve can then be plotted against the median permuted curves and the 95% interval of the permuted curves to demonstrate how the observed curve differs from the null.

16

Because none of the specifications are independent because all use some overlapping variables, traditional tests that assume independence are not appropriate. Instead, I will base decisions on whether curves differ from the null on three tests: (1) the median overall point estimate within each specification curve, (2) the percentage of specifications that are of the dominant sign, and (3) the percentage of specifications with the dominant sign that are also significant. Each of these results in p value constructed from the number of permutations that meet each criterion divided by the number of permutations.

Transformations   Transformations have all been documented in the codebook.xlsx file included in this preregistration. In general, transformations were done to make categorical variables across studies have similar codings to aid interpretation of their effects.

Inference criteria   In study 1, which uses propensity score matching and bayesian MLM, inferences will be based on the Bayesian uncertainty interval, which indicates the 89% interval of samples from the Bayesian model. The critical test will be whether this interval (of log odds) overlaps with 0.

In study 2, which uses specification curve analysis, inferences will be based on three criterion through a permutation based procedure: (1) the median overall point estimate within each specification curve, (2) the percentage of specifications that are of the dominant sign, and (3) the percentage of specifications with the dominant sign that are also significant. Each of these results in p value constructed from the number of permutations that meet each criterion divided by the number of permutations.

Data exclusion   Where possible, all data will be included, meaning that different participants will be in different models, depending on whether they provided personality, outcome, covariate, and moderator variables for a specific combinations.

Missing data   I will use multiple imputation within each study to get complete data for propensity score matching in study 1.

In study 2, I will use raw data and maximum likelihood estimation within `lme4`.