Sample selection for victimization immunology study

Justin Chumbley

Abstract

Exactly 500 people were invited to participate. Each person belonged to one of four victimization exposure groups (excluded, physical, neither or both) and to (at least one) quadruplet of other participants with whom they had different exposure but exactly matched gender, teacher rated bullying, aggression, anxiety and competent conflict resolution. The maximal subset of these 500 that can be 1:1 matched is 348 = 4 groups of 87. Victimization was defined as at least 3 times in the last year on at least one wave.

As an aside, we relaxed the requirement of one-to-one matching at the first invitation, i.e. permitted some participants to have multiple matches, to ward against loosing subjects due to non-response. The 500 themselves – the most that Z-prozo were willing to invite - are themselves a subset of 828 subjects (= 417 + 188 + 99 + 124 across the four exposure groups) who can be matched. This would be the sample size had we pursued the most exhaustive, many-to-one matching design (each victim is matched to as many control subjects as exist).

Background

This memo summarizes our work on research design: how to strengthen causal inferences about victimization and gene expression. Our guiding substantive concern has been to identify the effects of social exclusion and physical victimization (from about ages 11 to 17) on gene expression (at age 20) in a causal framework. Throughout, we aspire to the ideal of non-parametric causal contrasts under non-parametric corrections for confounding, e.g. simple differences of mean among matched groups. We will be unable to completely achieve this goal to every reader's satisfaction. Indeed, given the data-scarcity, one proposal has been to entirely postpone consideration of bias-correction from the sampling design phase to the analysis phase, i.e. eschewing *ex ante* matching for *ex post* parametric adjustments. This would obviously increase the sample size – and therefore power – at the expense of weaker control of bias (correction would now depend on correct parametric specification). The basic issue is a bias-variance tradeoff. In general, causal methods spend degrees of freedom, precision or power to correct (reduce) bias[1].

It will become clear that, if we wish our final outcome analysis to be based on non-parametric matching, it would be foolish to use a random sampling approach: A random sample of non-victims will obviously not be matched with

[1] For example, as we see below, matching corrections essentially prune subjects, accepting higher variance in return for reducing any bias from the matching variables. (Although "coarsened" matching then recovers some precision by relaxing control of confounder bias). Alternatively, we may exchange our non-parametric methods, which correct bias in general, for parametric methods – which only correct bias if well specified. This again buys subjects and precision, at the cost of possible new misspecification bias.

victims on key pretreatment variables. Rather, to achieve this matching, we would have to exclude or prune some subjects, subjects that we have already invested substantially in measuring. Conversely, if we do not plan to use non-parametric matching in the final analysis, but are content for our final analysis to rely entirely on parametric corrections, then random sampling is a less bad idea.

One theme below is how to compress the many-to-many relationship between many victimization items and times on one side, to many genes on the other. The coarsest solution is to compress each side of the equation to a single number per subject (e.g. mean CTRA versus binary victimization status). At the other extreme one would quantify the mutual information between these two vectors, via multivariate canonical correlation methods and similar.

Causal considerations

We do not aim to estimate or test properties of z-prozo's original target population as a whole, nor to predict the outcome of new subjects from that population[2], nor to describe observable properties of the collected sample. Instead, we aim to predict what would have happened to victims in our sample, had they (counterfactually) not been victimized. This goal is motivated by the standard **definition** of causation, as a difference of actual and counterfactual outcomes. We consider 3 cases below.

1. Suppose you believe that **nonvictims' CTRA** corresponded perfectly to **victims' counterfactual CTRA**, had they not been victimized. With this judgment of exchangeability you directly reach the desired counterfactual: the **average causal effect of victimization** is then simply the difference in average CTRA across the two groups. Thus association equals causation without needing corrections, and the sample size for causal comparisons is maximal. This enables us to assess nuanced contrasts, such as between different temporal patterns. These include recency effects, and more generally associating different CTRA outcomes with different trajectories.

2. Now suppose you believe that victims differ on an observed set of extraneous[3], time-invariant cause of CTRA[4], but **do not differ** on any time-varying cause of CTRA[5]. Then the causal inference method of the last paragraph is only valid **within fixed levels of this confounding factor** (where variation due to the confound is clamped). Popular tools for correction – regression, stratification, matching, … - assume this *conditional* exchangeability of counterfactual outcomes. Section 2 proceeds with this assumption: it supposes various scenarios – different variable sets a reader may deem sufficient to correct counterfactual predictions. For each, it calculates the sample size – and hence the

[2] Had we been concerned with prediction only, there would be no need to precisely specify any prediction variable, nor assess its role independently of other variables: a black box approach- mapping the whole z-proso data-set to gene expression would suffice.
[3] Not on the causal path from victimization to CTRA.
[4] For example, victims may differ in time-invariant ethnicity distribution, which in turn implies spurious non-causal CTRA group differences.
[5] An extraneous factor whose value can freely evolve in the time between the initial victimization and outcome. For example, moving to a new school may increase exposure to both victimization and environmental pathogens which impact CTRA.

precision - of unbiased causal contrasts between matched groups (coarsened exact matching), unbiased for any reader that accepts the premise of the scenario.

3. Finally, suppose you believe **time-varying victimization is associated with other, extraneous time-varying causes of CTRA (what are they?)**. Then no method so far discussed here can generally correct confounding. Graphs – e.g. single world intervention graphs or SWIGS championed in causal epidemiology - offer a way to derive whether correction is possible by any method.

Keep in mind below that every causal argument about the effect of a treatment on some outcome requires a premise about the determinants of the treatment, "the treatment assignment mechanism". In general this means asserting some restrictions on the treatment assignment mechanism (e.g. no unmeasured confounders). Arguments are not causal, unless they make an assertion about the mechanism governing treatment. Graphs are a simple way to express qualitative assumptions about treatment assignment mechanism. For claims about the effect of a time-varying treatment, the assignment mechanism is obviously a dynamic process, and obviously not a simple Markov process, e.g. the simple LTA model (below). Modern work in formal epidemiology – e.g. James Robins - has clarified this.

As seen below, our favored solution will be to use a non-parametric matching procedure for the "pre-treatment" period and then G estimation for the treatment period, if and when concrete concerns about time-varying confounds arise.

Section 1

Definition of Treatment

This section defers any concerns about confounding and bias, in order to focus on the question of treatment definition. We have **<u>underlined</u>** the most consequential choices below.

How is adult CTRA influenced by historical experience on our four victimization dimensions? Does it depend only on recent social exclusion? Or is it sensitive to other recent social stressors? How does it depend on historical experiences? Is a multi-year pattern of chronic exclusion needed to activate CTRA, or is a single year of exclusion sufficient, or a combination of different types of victimization for a single year necessary? To answer all these questions, while "letting the data speak for itself", imagine that we have "big data": enough subjects to precisely estimate the expected outcome under each treatment pattern (e.g. N subjects in each of the $(2^4)^4 = 65536$ possible exposure patterns for 4, binarized outcomes over the 4 waves). Then, the outcome difference for every pair of treatment patterns identifies the corresponding causal effect contrast (yielding 2,147,450,880 treatment effects!). In fact there are far fewer subjects than possible patterns. This exercise illustrates why we must define exposure groups more broadly, while trying to retain causally important distinctions about the

nature and timing of exposure. Any treatment should not be defined so narrowly that few qualify, nor so broadly that our conclusions loose resolution because victims differ in important ways.

A question then faced by all solutions below is to define the **type of victimization**. We propose to focus on social exclusion and physical victimization for several reasons (the two other items are insulted/mocked and robbed/had stuff damaged). First, there is a well-established and long-standing interest in these types. Second, we wish to study these types as opposed to a composite measure of any form of victimization.

Despite this focus, we still face a choice about what to do with the other two "nuisance" items. If we entirely ignore them, we neglect that they may correlate with the items of interest (and hence actually explain outcome differences). If we stratify on them, we attain smaller sample sizes at the design phase and risk conditioning on post-treatment mediators (if, for example, being insulted mediates the effect of early exclusion and outcome). The tradeoff is unavoidable and should be based on expert domain knowledge. The same tradeoff arises in our choice of whether to ignore victims' **perpetration status** or stratify on it. Again the right tradeoff is personal: it depends on your domain knowledge (it can be derived from a causal graph of your domain knowledge).

We now discuss some concrete solutions to the problem of defining treatment. We currently favor solution 1.

Solution 1: Thresholding

Apply an (arbitrary) threshold over all time (i.e., a person is classified a victim if she/he crosses the threshold **at any measurement occasion**). This induces a partition of subjects into victims/non-victims. This definition permits diverse temporal exposure patterns to enter the same broad class of "victims". Generally, the challenge for non-parametric matching efforts is to attain sufficiently high count per treatment category or cell. This count varies depending on the following parameters:

> **a. Cut-off for what constitutes victimization (e.g., 2 =1-2 times, 3 = 3-10 times/year);**
>
> **b. Number of waves of data used (3 or 4 possible; ages 11, 13, 15, 17);**

Ultimately, in the following section, we will see that this count also varies (decreases) with the number of pretreatment variables that we require be matched across victim and non-victim groups.

Solution 1 does conflate (perniciously?) subjects who exceeded threshold long ago versus recently. Conversely, it has the advantage that treatment groups

depend on the simple parameters above (and not the parameters of a more speculative model, as in solution 4 below).

Solution 2: Cluster subjects by the criterion that some cumulative "dose score" exceeds threshold. This is akin to solution 1: victims defined according to solution 1 will have higher average dose scores. It has the disadvantage that dose scores are meaningless on close inspection (you cannot add or average the ordered categorical response data from the victimization questionnaire, because they are not interval or ratio data). It inherits the disadvantages of solution 1, the threshold is arbitrary, as is the choice about how to calculate the dose (over years, over items, over both?). We nonetheless briefly present some descriptives below.

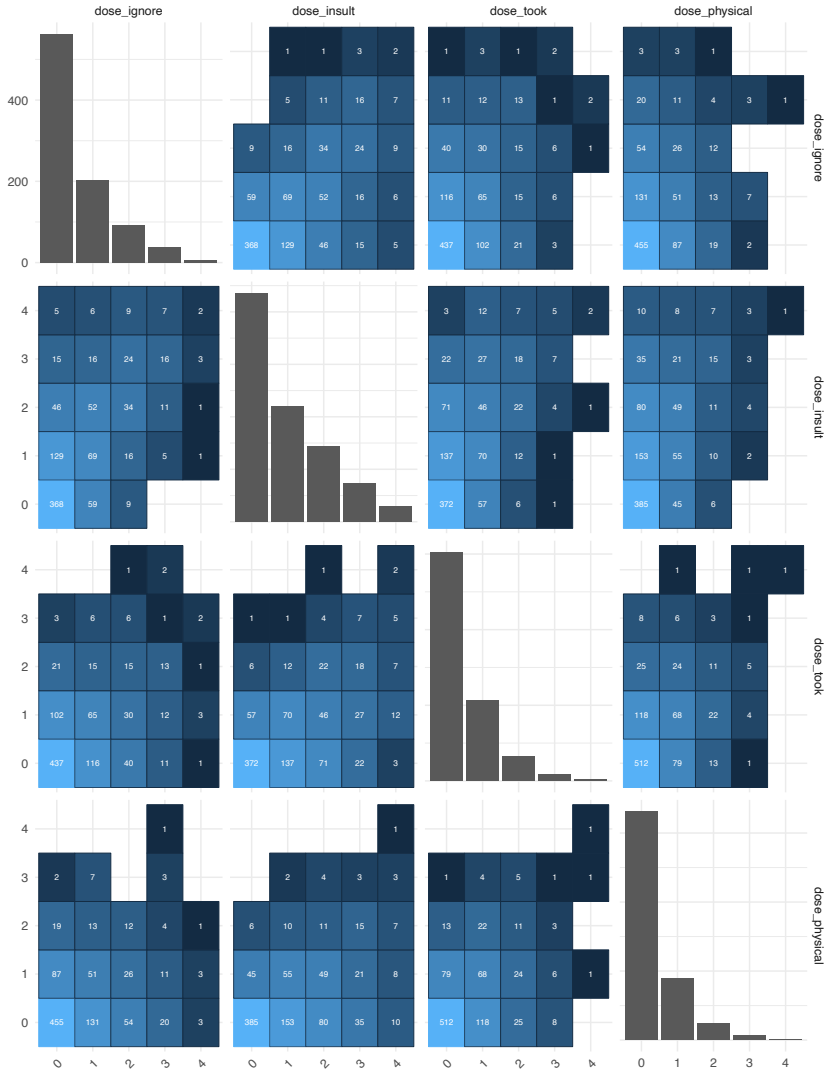


Figure: A cumulative dose score, over time and for each item, was calculated by first binarizing item frequencies (1 if and only if the frequency of this item of victimization exceeded 2 times a year), and then adding these over 4 waves. Thus the range of this score is 0 to 4. This table cross plots the dose score on different items.

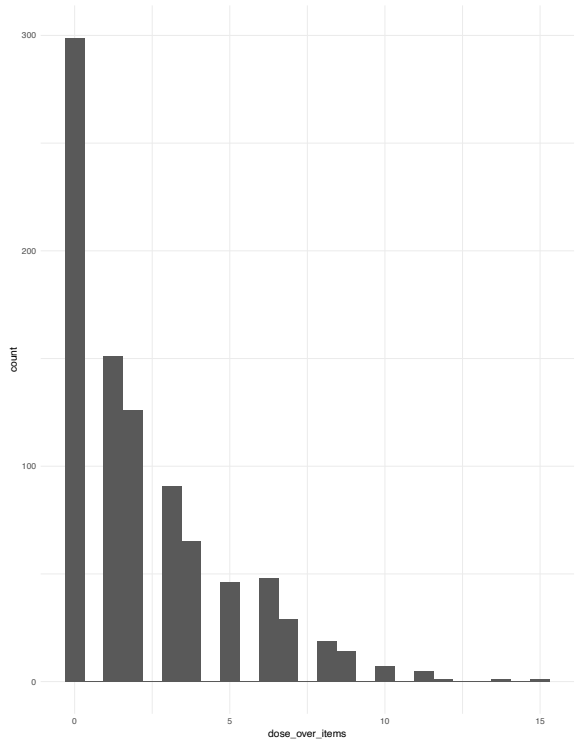


Figure: Both time- and item-cumulative dose score. We here have further summed victimization dummies over items, as well as waves (previous figure), to give the least interpretable score. Thus it is unclear whether a high score reflects multiple items per year or one item for multiple years.

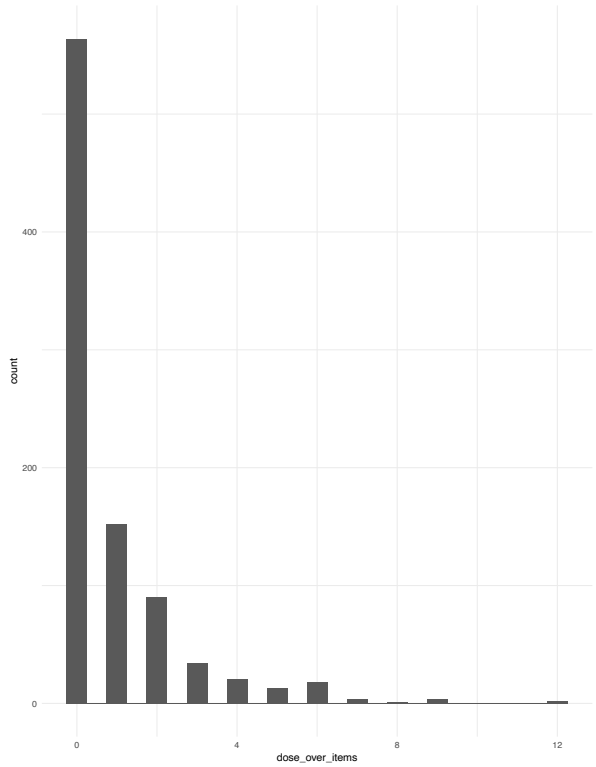


Figure: Same total cumulative dose scores, but with a more conservative definition of victimization in which individuals must have more than 10 events/year of a type/item of victimization before being classified as victimized on that item/year.

Solution 3: ignore subjects with rare trajectories. There are 362 distinct observed victimization patterns in z-prozo (assuming binarized bullying frequency > 2 times per year, on all 4 of the bully items), but only 16 patterns with > 5 subjects, which include a total of 470 subjects. This is about half of the 903 subjects for which we have complete exposure data. These "high-frequency" patterns have the following sample size: 299, 21, 20, 14, 13, 13, 12, 11, 10, 10, 10, 9, 8, 7, 7, 6. The most prevalent pattern being "never victimized on any item", at 299. We can repeat this exercise with our two focal items, exclusion victimization and physical victimization. If we restrict to the social exclusion and physical items, *ignoring property and insulting items*, these numbers are a little larger. If we restrict to the social exclusion and physical items, *while requiring property and insulting items are sub-threshold (see above)*, the numbers are smaller.

Solution 4: adopt a parametric restriction on the effective number of treatment patterns.

A latent variable dynamic model permits us to cluster subjects on the assumption that different observed patterns correspond to the same latent pattern (superficial differences due to "measurement error"). This reduces the number of distinct treatment trajectories if, as usual, the latent space is smaller than the observer response space. In addition, the mean outcome of subjects assigned to two substantively distinct latent trajectories of interest could then be qualitatively contrasted, in principle.

LTA seems to be popular, but has oddities. Being Markovian, it assumes that the treatment assignment mechanism lacks memory: victimization at time *t*, cannot depend on victimization at time *t-2*, except via victimization at *t-1*. This restriction does not cohere with the belief, for example, that both perpetrators and the victim condition their behavior on long-term memory of prior victimization, or that early victimization effects are mediated by long-range effects. LTA in particular, also assumes discrete latent classes, and that the number and type of classes is time-invariant. It models the fraction of state transitions at each time, but not the fraction of individuals following a sequence of transitions. It is not an explicitly causal model of victimization: it does not model counterfactual transitions, i.e. what would my transition have been, had I been in a different latent state. If the goal is simply to identify "similar" treatments, from the data itself, any clustering algorithm will do.

All such approaches carry the possible disadvantage of being complicated: the definition of treatment groups is sensitive to the specification of a parametric model. In addition, any probabilistically sound clustering of subjects into latent victimization trajectories only yields subject-specific posterior distribution over such trajectories: some further choices are needed to choose a single latent trajectory (e.g. maximum *a posteriori* trajectory "MAT" is needed). Further, even if model selection methods yield a model with low state-space dimension, any qualitative comparison between two groups will require further choices about

how to group subjects with “similar” MATs (in order to get enough subjects per group).

The next figure presents a crude clustering of subjects trajectories for ignore and physical items, ignoring the other two items (We assumed there were 15 clusters, but only report clusters containing more than 10 victims, where victim defined as >2 times a year). The subsequent figure then presents a similar partition of subjects who were victim of ignore/physical, *among those subjects who do not report any victimization on the other two victimization items*. It will be seen that very few subjects remain after this restriction is imposed. This reflects that items are somewhat correlated.

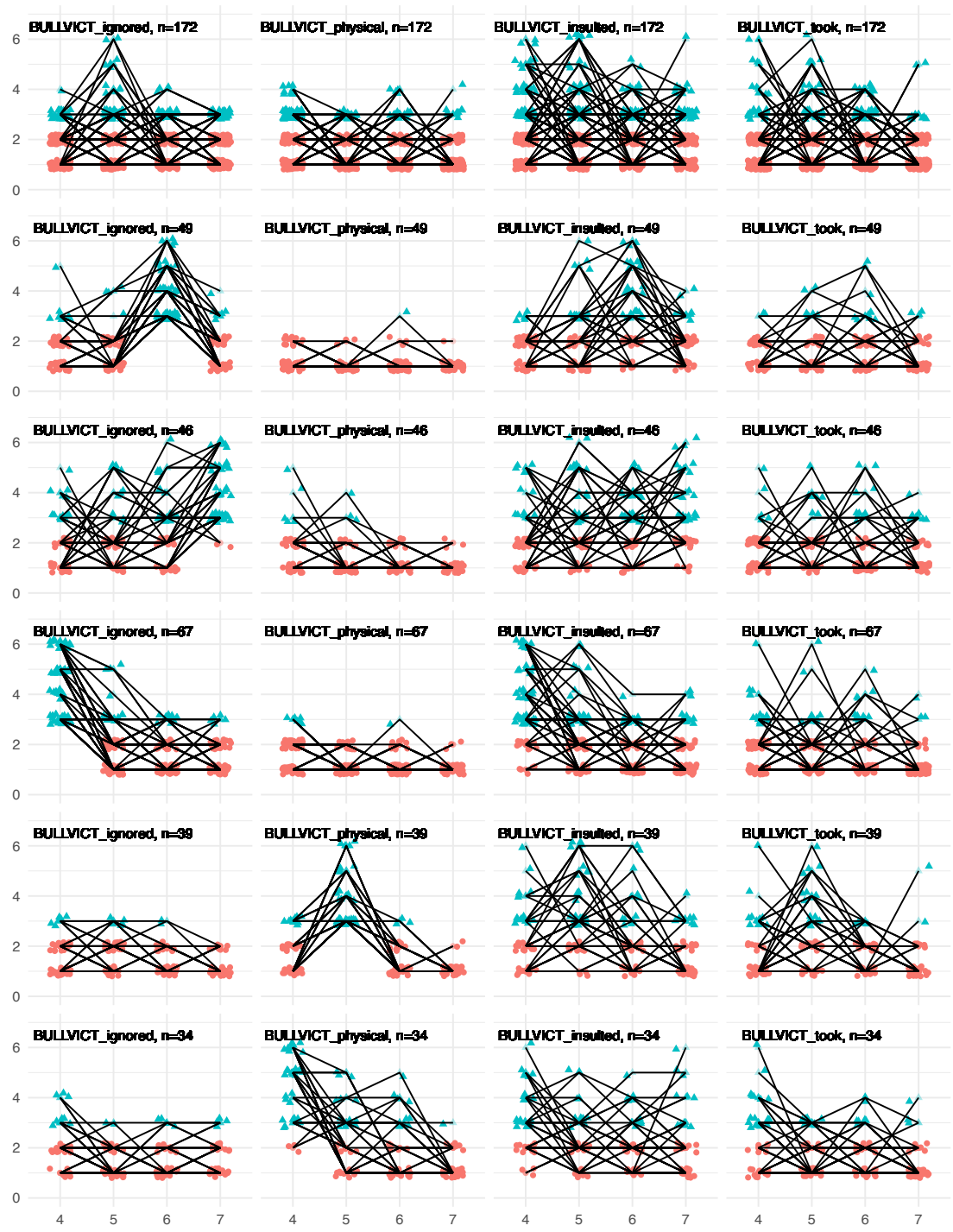


Figure: Cluster 2 items of interest, while ignoring insulted and property items.

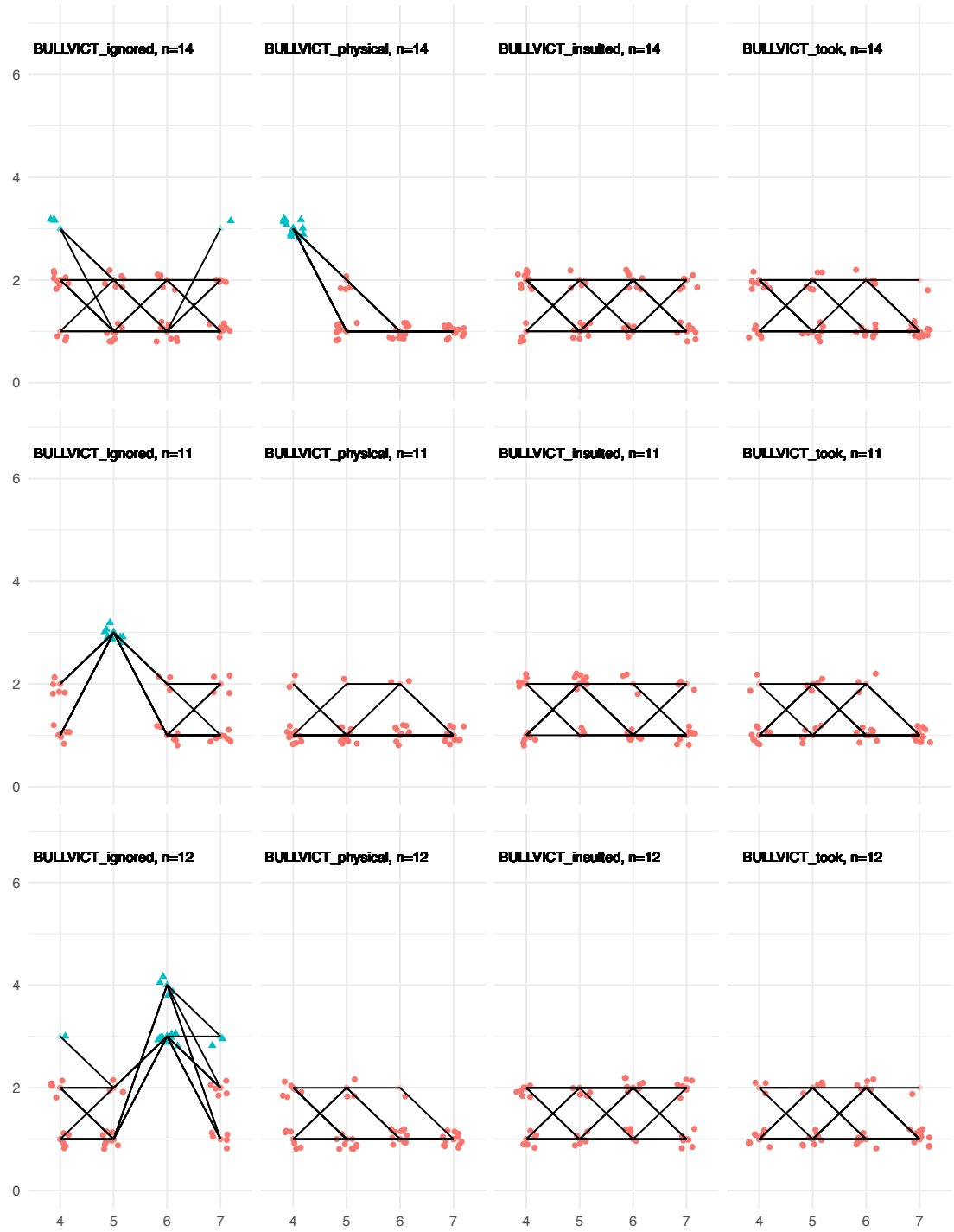


Figure: Same but now conditional on never insulted or victimized.

Solution 5: Random sampling

The above solutions involve non-random sampling from z-proso. Randomization would facilitate inferences about Z-proso's target population, but this is not our goal (see "Causal considerations"). We seek an optimal sample, not a random sample. Non-random sampling is common, as when a matching procedure prunes control subjects from some super-population on a systematic basis, in order to fortify causal claims (see next section). Picking subjects at random

would result in too few (rare) victims, although stratified sampling to oversample victims solves this problem.

A specious argument for random sampling is that it avoids clear *ex ante* definitions of treatment, outcome and confounder. The gold standard for causal inference – randomized experimental trials – require pre-committing to a treatment and primary outcome. Observational studies, lacking any experimental intervention, do not require a lower burden of proof.

Section 2

Correcting for time-invariant, "pretreatment" confounds

Here we discuss non-parametric adjustments, namely matching, for time-invariant confounding of the relation between victimization *as defined by solution 1* above, and CTRA outcome. See "Case 2" of "Causal considerations" section above.

Solution 1, again

We now revisit our preferred design (solution 1), while now admitting that pretreatment variables may confound counterfactual prediction. We have pragmatically considered different sets of possible pretreatment confounds. For each we implement a non-parametric matching procedure. This procedure ensures that our sample has the following property: for every person in one treatment, (at least) one control person exists *in every other treatment group who are exactly matched on all of the named pretreatment confounds*. Formally, this design then satisfies the necessary condition of *positivity*. *Due to the unknown level of non-response from our upcoming sample, we cannot at this time guarantee that positivity will still hold in the sample who consent to give CTRA data.*

In summary, the main goal here is to report the attainable sample size for a matched comparison between victimized and non-victimized individuals. This sample size decreases with the number of matching variables. Matching was done on (all subsets of) the following list of pretreatment variables: (a) Gender, corporal punishment ("b" in the table and figures below), (c) teacher rated bullying, children, (d) above median aggression, (e) above median anxiety, (f) any developmental disorder, (g) above median ADHD, self-reported bully victimization,(h) percentile group of social locus of control, (i) self-control, (j) violent peers, (k) lack competent conflict resolution, (l) life events, (m) Percentile Group of SES, (n) parent immigration. Gender was matched in every case analysis.

In all graphs below, the timing of victimization has not been specified: we consider short trajectories of 3-4 waves (see Figure 1 below), namely waves 5-7 and 4-7 respectively. You are a victim if you exceed threshold at any time in this trajectory. The thresholds we entertain are **1** = never, **2** =1-2 times, **3** = 3-10 times/year. These are indicated by "thres" in the Figures.

Design 1:

We start with the weakest definition of victimization: *Ever victimized on any item*. This definition ignores whether the victim also has reported (post-treatment) perpetration behavior. The following four graphs present the best options.

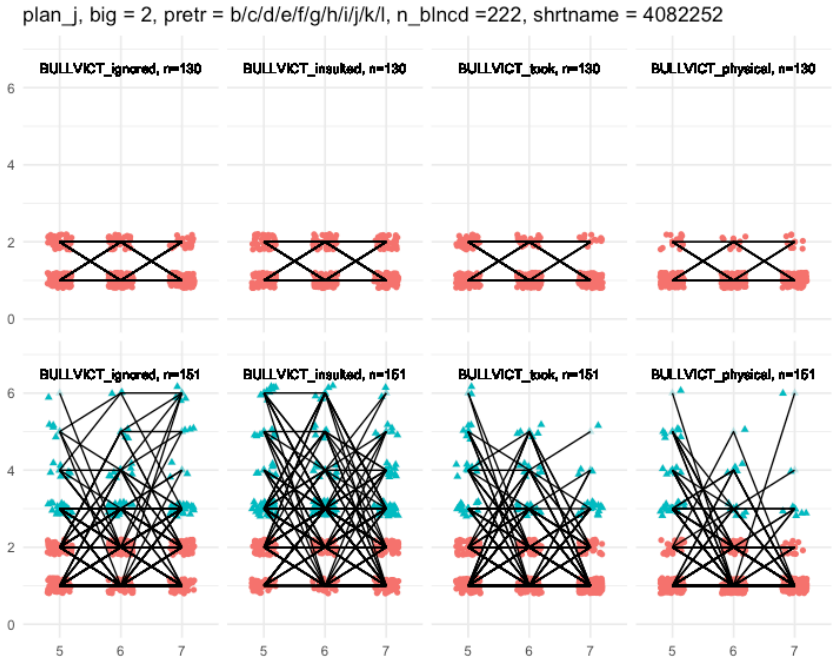


Figure: We control for 12 pretreatment variables (including gender).

Design 2:

Next we consider the same, weak, definition of victimization. However, we exclude any subjects from our analysis that reported perpetration on any item, at any wave. Thus we consider "pure victims". Note, this restriction means we cannot control as many pretreatment confounds and still achieve such high sample sizes. Note that this design is fatally flawed for causal inference if we believe that post-treatment (post-victimization) perpetration mediates the effect of victimization on CTRA outcome. In this case we should not condition, i.e. stratify, on the value of this variable.

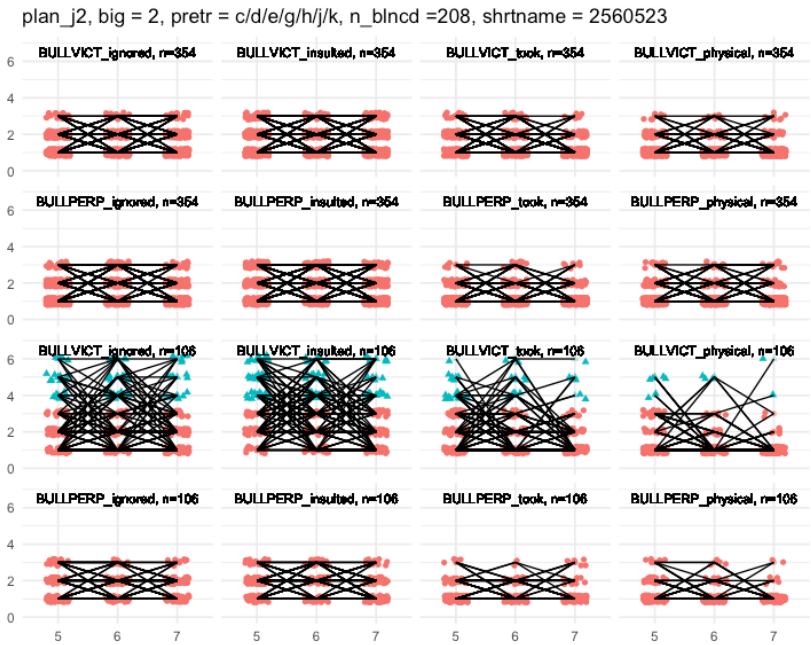


Figure: Four waves, threshold of > 10 events/year.

Design 3:

Next, we ignore items on property and insulting, and try to isolate exclusion from physical victimization. This is a 2 x 2 design.

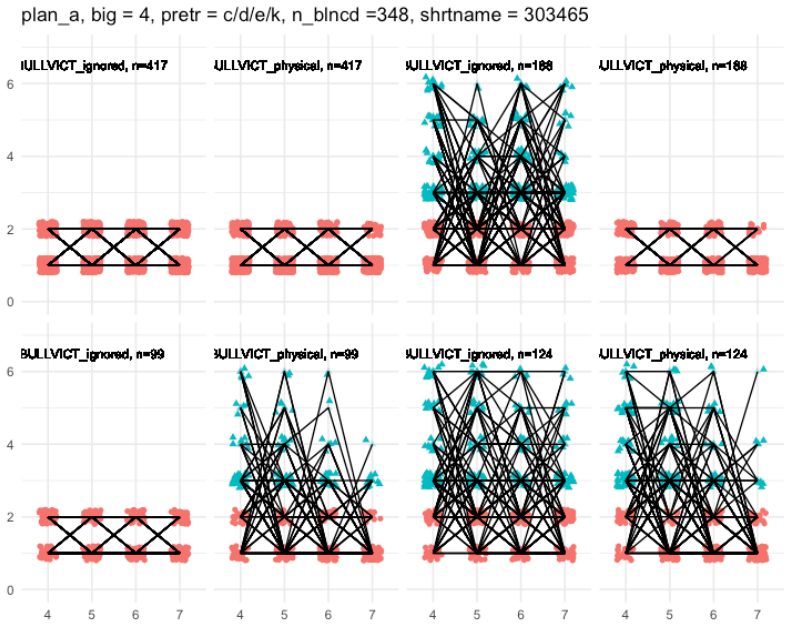


Figure 1: Top-row panels a/b (reading from the left) depict individuals neither excluded nor physically attacked. Top-row panels c/d depict individuals excluded, not physically attacked. Bottom e/f (from left) depict subjects selectively exposed to physical victimization. Bottom g/h are victimized on both dimensions. Other victimization and perpetration items have been ignored in the treatment definition, i.e. individuals in either group may or may not have been victims of insulting/robbing/destroying, and may or may not be perpetrators. These two groups are matched on 4 pretreatment variables (the three given, plus

gender, for which we always match in every design presented here). If we pursue many-to-one matching (each victim is matched to as many control subjects as exist) then the sample size for this example is 417+188+99+124. This is seen by summing the *n* in each cell of the 4 cells in figure. If we pursue one-to-one matching - within each strata, an equal number of controls and victims – then sample size for this design is the parameter „n_blnced“, here n_blnced = 348).

Note that we can also consider a subset of this design. While this breaks the factorial symmetry, and therefore may seem unappealing. It does however permit us to match on more variables for the remaining cells. The next figure, gives an example.

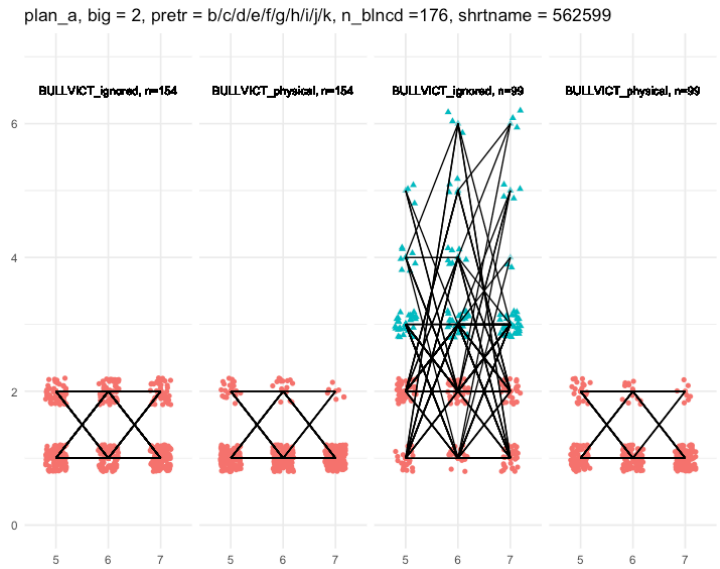


Figure 1: Panels a/b (reading from the left) depict individuals neither excluded nor physically attacked. Panels c/d depict individuals excluded, not physically attacked. Other victimization and perpetration items have been ignored in the treatment definition, i.e. individuals in either group may or may not have been victims of insulting/robbing/destroying, and may or may not be perpetrators. These two groups are matched on 11 pretreatment variables. If we pursue many-to-one matching (each victim is matched to as many control subjects as exist) then the sample size for this example is 253=154 +99. This is seen by summing the *n* in the left half, and right half of this figure. If we pursue one-to-one matching - within each strata, an equal number of controls and victims – then sample size for this design is the parameter „n_blnced“, here n_blnced = 176).

Design 4:

We now assume that we have decided to execute one version of the previous „Design 3“ above. Namely, the matched design with 348 subjects in total (4 groups of 87) depicted below. This group is one-to-one matched on gender, teacher rated bullying, aggression, anxiety and competent conflict resolution.

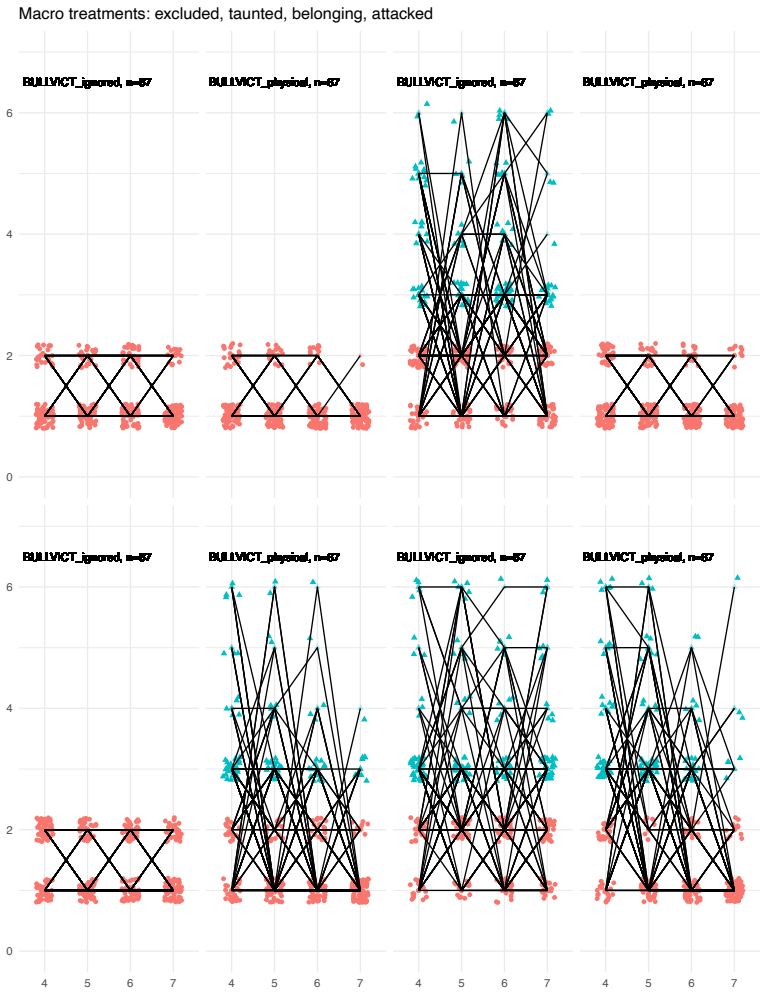


Figure: A one-to-one matched design.

In what follows, we discuss whether secondary questions may be confidently answered from this same set of 348 subjects. Accordingly, the next plot depicts a secondary design, which „reuses“this 348 sample to ask the question we had originally considered in „Design 1“ above, namely: what is the effect of victimization under it's weakest definition (under a weak definition of victimization any item at any time). It is apparent that we have 56 non-victims and 276 victims available to answer this question, after matching for the same pretreatment variables in our seeding design – i.e. gender, teacher rated bullying, aggression, anxiety and competent conflict resolution. 332 = 56 + 276 is less than our original sample size of 348. In general, a second iteration of matching will again prune subjects. For this reason, we will sometimes below give the sample size for a secondary design that omits any matching procedure. This is done on the understanding that parametric corrections are the only realistic option for control of pretreament confounds. In general, even if no subjects are lost to pruning, the design size for secondary questions may still decrease if there is missing data on any new variable used to define the new comparison groups of our secondary design– e.g. insulted and robbed in this case.

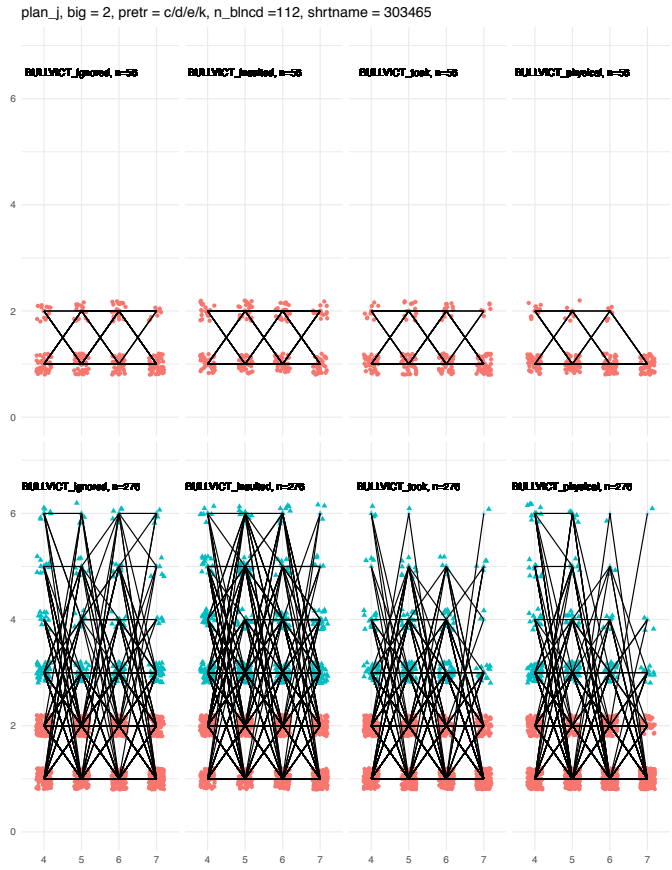
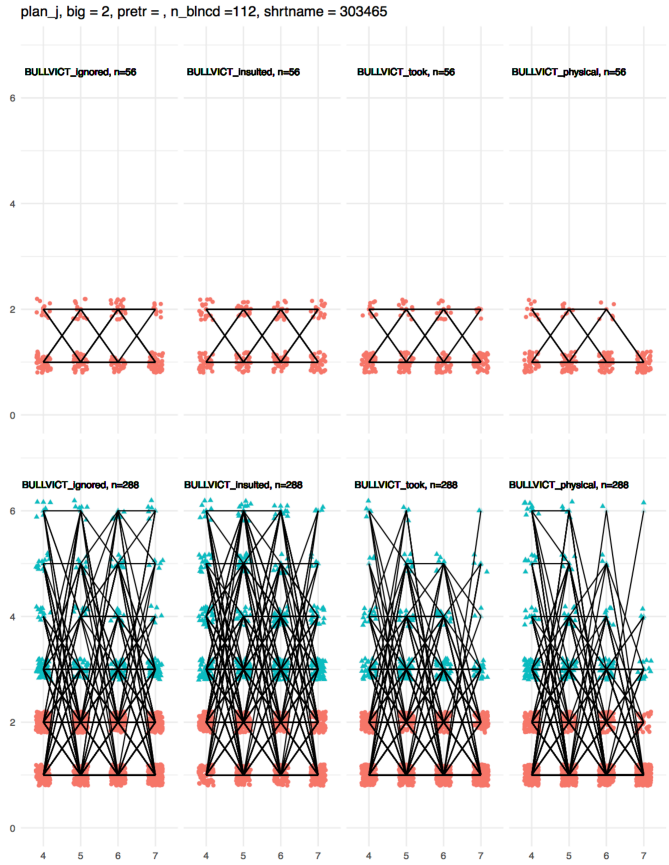


Figure: Two new comparison groups, chosen to answer our secondary question about the effect of victimization in it's weakest sense. These comparison groups are subsets of our original, primary sample.

The next figure is very similar, but ommits the matching step from our secondary design and therefore has a slightly higher sample size 344 = 56+288.



The following design asks the same question as our primary design – i.e. the selective effects of social exclusion and physical punishment – but on a different subgroup of subjects. In particular, it assesses these effects among „pure victims“ of social exclusion (a subset of the 348 seeding sample). These individuals are, by definition, not perpetrators on either exclusion or physical dimensions (hence „pure“ victim). Note that this restriction is based on a post-treatment variable (wave 4-7 perpetration behaviour) and is therefore inherently dangerous, if – for example - we believe the early effect of victimization might be mediated by later perpetration (see g-methods).

The nuissance victimization items have been ignored. There are four cells in this design. The sample size here is smaller, because we restrict to a subset of subjects (who were not perpetrators). The design – depicted below – is many-to-one matching design, in that includes all subjects who have at least one comparison individual, alike on all pretreatment control variables, in all other conditions. (It does not insist on one-to-one matching in the secondary design). The sample size is 97, but there are not many subjects per group.

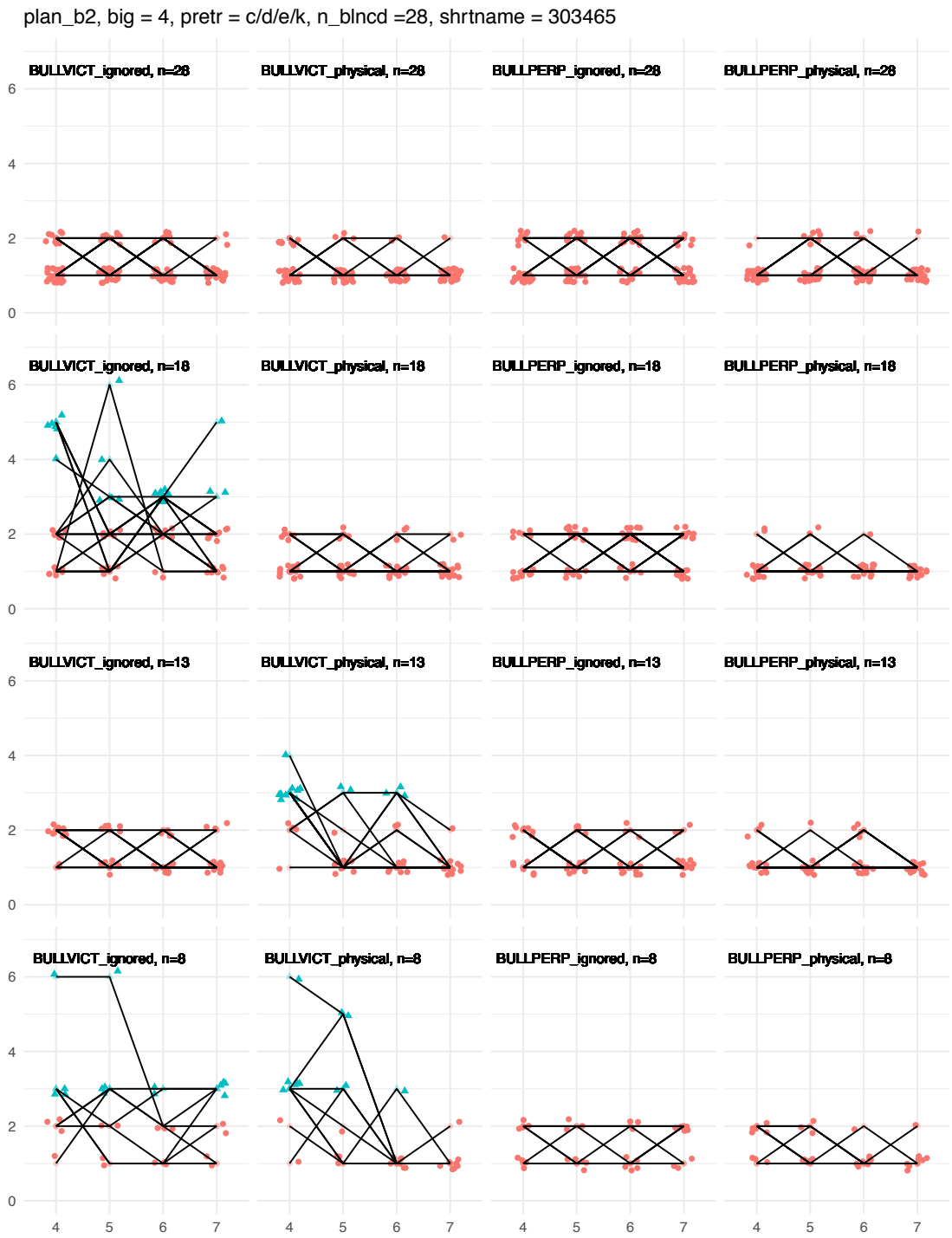


Figure: Comparison of the social exclusion and physical attack amoung „pure victims“. Each row depicts one comparison group.

The sample sizes here are negligible. If we relax the requirement that subjects are matched on the same variables as in the primary design, we attain the following.

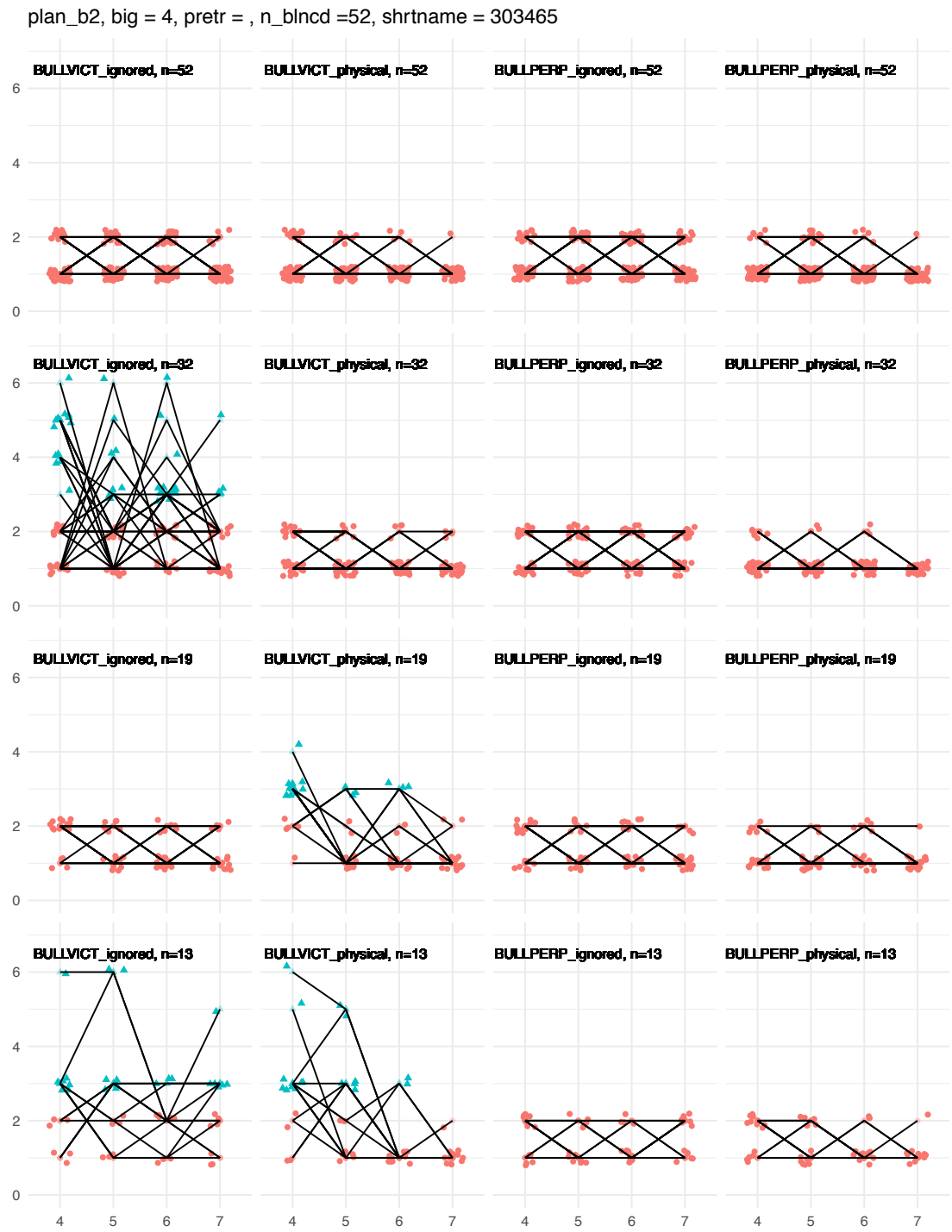


Figure: Our primary design, but on the subset of our primary sample who manifest no perpetration: neither physical or social exclusion.

This design has n = 115, with groups of size 13, 19, 32, 52. This shows that we do not have much power to examine our primary effects *in the subpopulation of pure victims.* This suggests that parametric methods may be required.

We now examine the power to assess (sociometric) network treatments of subjects within our primary sample. For details, see the next section. Under definition „i1“ of peer exclusion (next section) we find that 130 subjects are available for a one-to-one matched design, which compares individuals who

were and were not excluded *by peer report*. If we use a many-to-one matching process, this yeilds good group sizes of 136 excluded and 212 not excluded. We present a summary of the in-sample sociometrics below. We conclude that sociometric comparisons are well-powered in our primary sample.

For reference please compare this to the same summary over the entire z-proso sample (not just our primary sample).

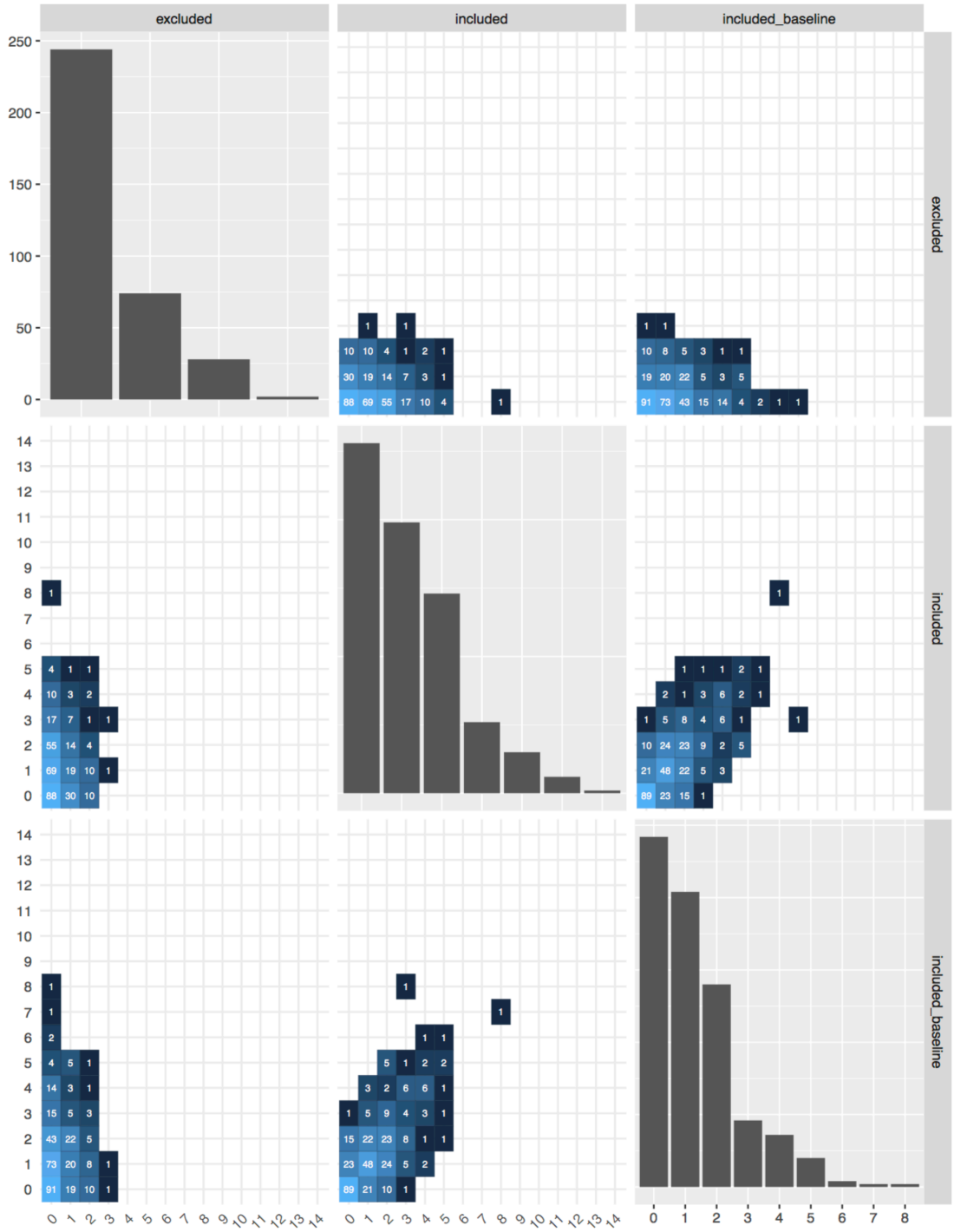


Figure: Some sociometric summaries with the primary sample.

We next consider variation in *atypical emotional attribution* among our primary sample. In brief, we created a score that measured how eccentric was each subject's assessment of the facial emotion in the standardized face set used by Z-proso, relative to their peers. A higher score means they are more atypical. We see that there is good variation in the scores for anger attribution in our sample. This provides hope that one may correlate this variation with adult measures of hostile face attribution, behaviour or fMRI. There is no ready causal question here, rather a simple measure of long range prediction.

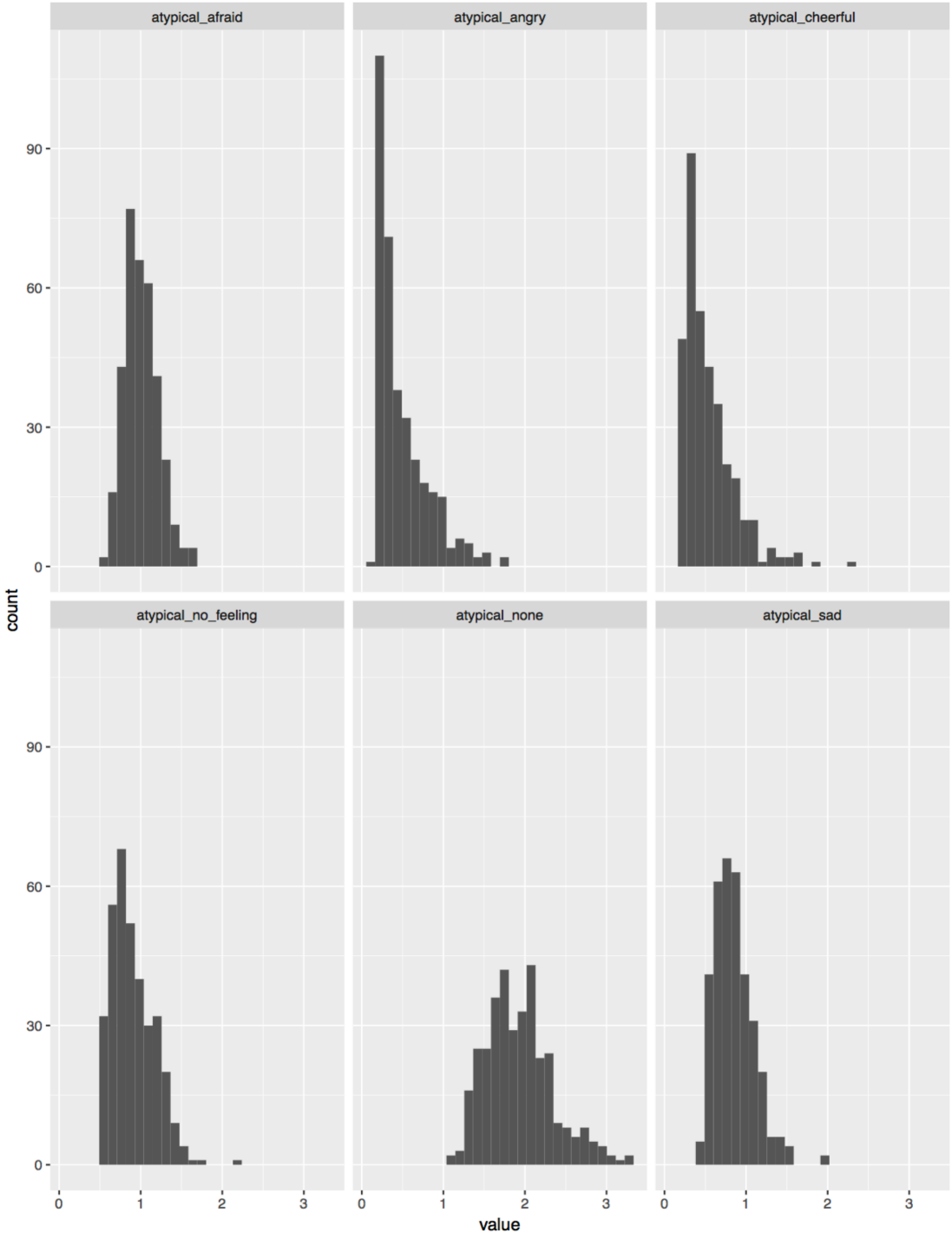


Figure: Face attribution atypicality.

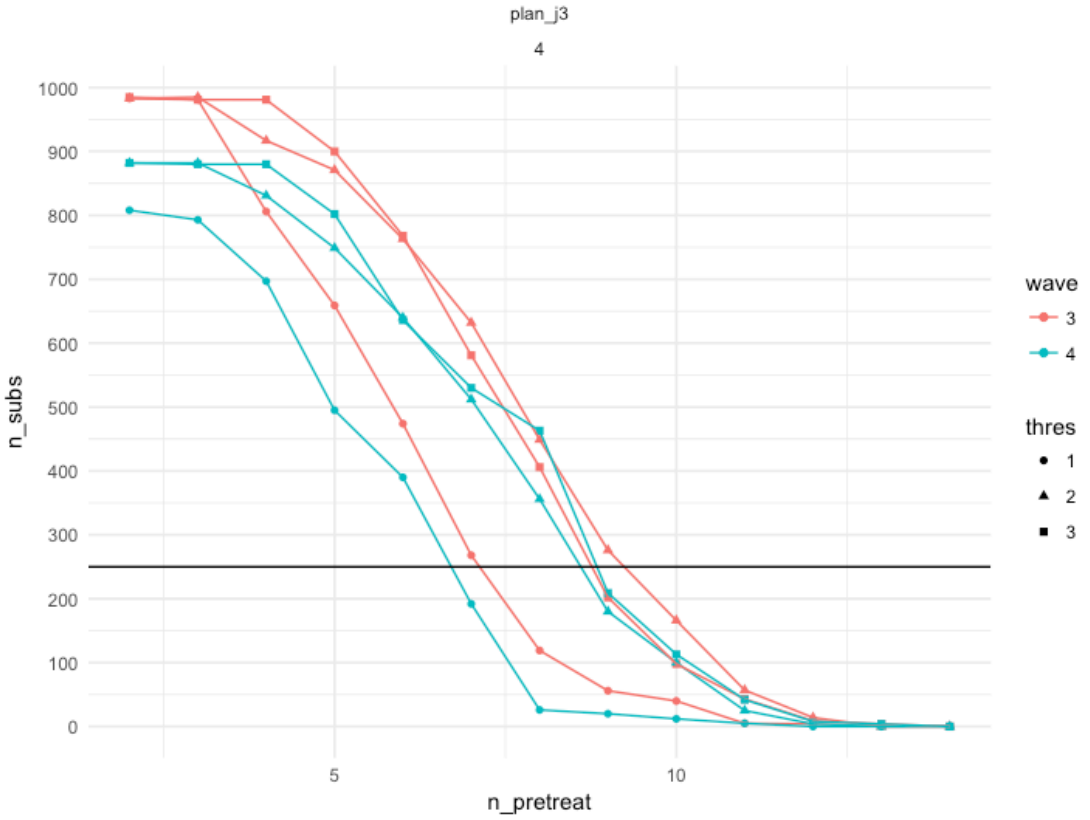


Figure: The x-axis is the number of pretreatment variables used for matching. The y-axis is the largest attainable sample size for the following 2 x 2 design: any victimization item at any time (yes/no) x any perpetration item at any time (yes/no). This design comes with the same health warning, because it stratifies on a potential mediator of the treatment effect.

This discussion of the sampling process has side-stepped the issue of time-dependent confounding which cannot, in my view be addressed at the sampling stage (c.f. analysis). Sample, like stratification and regression adjustment, is inherently a conditioning operation: selecting – that is conditioning – on the basis of post-treatment variables carries a hazard. Incidentally, this is one justification for ignoring perpetration status during the sampling (in contrast to common practise in the bullying literature of cross classifying bullys and victims).

Section 3: time-varying confounds

To reiterate, the above lists of pretreatment variables are intended to include all *time-invariant* confounds: pretreatment causes of victimization which also predict outcome. To identify the effect of *longitudinal victimization* we must also exclude *longitudinal confounds*. This is because spurious associations can be induced by any time-varying factor which both influences ongoing victimization and predicts outcome. Adjustments for longitudinal confounding require correct knowledge about its structure, which is non-trivial and subject to debate. To take one example, a chicken and egg problem emerges if dynamic peer victimization were actually a response to the child's evolving behavioral dispositions, such as their social perpetration or withdrawal, and vice versa (i.e. if there is treatment-confounder feedback)[6]. Critically, we are not generally permitted to stratify our sample on post treatment factors. For example, if all of the effect of early victimization is mediated by later perpetration, stratifying on perpetration will entirely eliminate the treatment effect. I believe that a safe solution is to avoid stratifying on the time-varying confounds – e.g. subjects' social withdrawal or perpetration status - and correct for this in the analysis phase. Of course such "correction" will rely strongly on a parametric model. Another solution is to blindly go ahead and stratify now on perpetration status. This seems appealing but is only valid if perpetration does not in fact mediate victimization effects.

Network definitions of early adverse social exclusion

This section changes tack completely and offers an alternative design, based on altogether different measures within the data set. This design has some advantages: it is not a time-varying or a self-reported measure of victimization, and it is very specific. It is reported here, because it might supplement or supplant one of the designs above.

In this section we first visualize the three socio-metric variables in z-prozo. These three variables quantify a) how many of my peers privately elected to include me on a bus trip, both in wave 1 baseline and wave 3, and b) how many of my peers expressly elected to exclude me.

The main goal is then to report the attainable sample size for a comparison between excluded/non-included individuals in wave 3 versus a matched sample. This sample size naturally decreases with the number of matching variables. Matching was done on (all subsets of) the following list: Gender, corporal punishment, teacher rated bullying "children, above median aggression, above median anxiety, above median ADHD, self-reported bully victimization, any developmental disorder, percentile group of social locus of control, self-control,

[6] Nevertheless, a causal effect of longitudinal victimization is identified by assuming that such dispositions and their causes are observed, and only affect the outcome *indirectly (*via their effect on future victimization). This assumption is violated, for example, if unobserved problems evolve at home which simultaneously influence the child's victimization at school – due to a child's social withdrawal or antisocial behavior at school - and impact the child's long run health *directly,* through a channel not mediated by victimization.

violent peers, lack competent conflict resolution, Percentile Group of SES, impt life events, parent immigration.

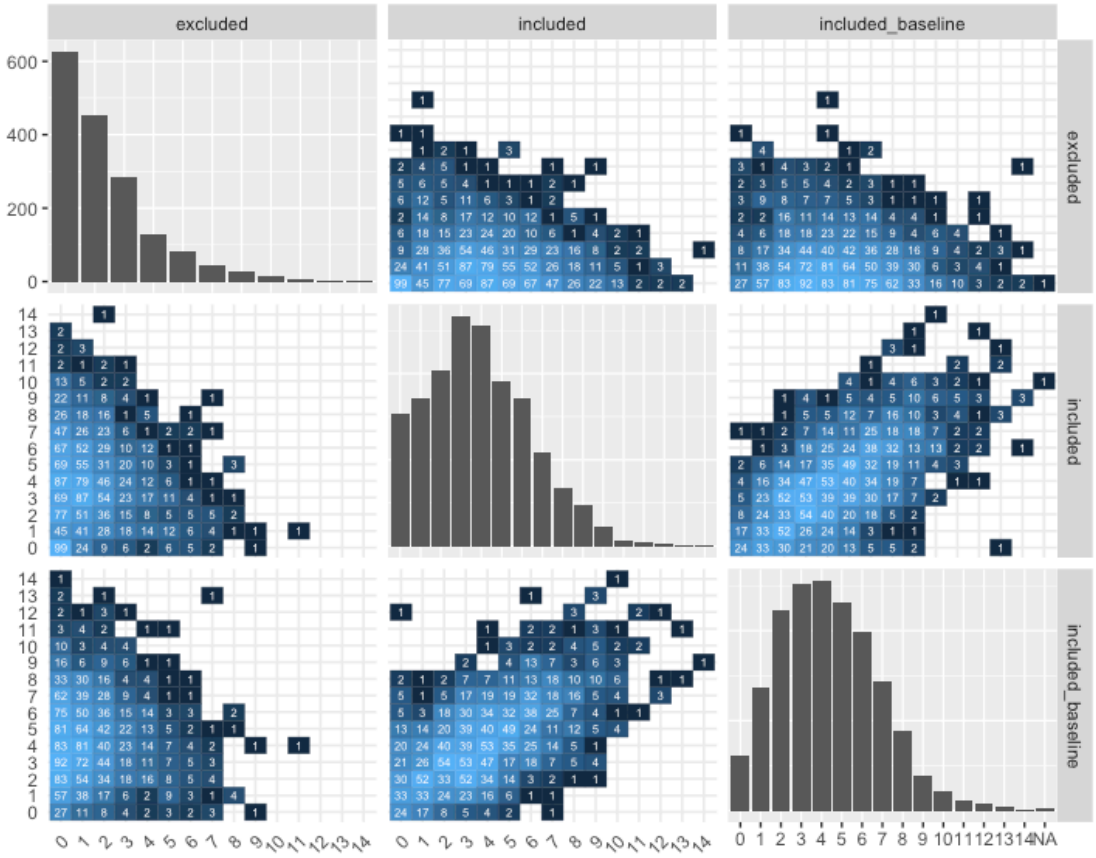


Figure: Marginal and bivariate distributions of inclusion/exclusion rate on each of the three socio-metric inclusion questions. The number in each cell is the number of subjects in that joint value. For example, middle-right betrays a positive correlation between the number of people who include me on wave 1 (baseline) and the number who include me in wave 3.

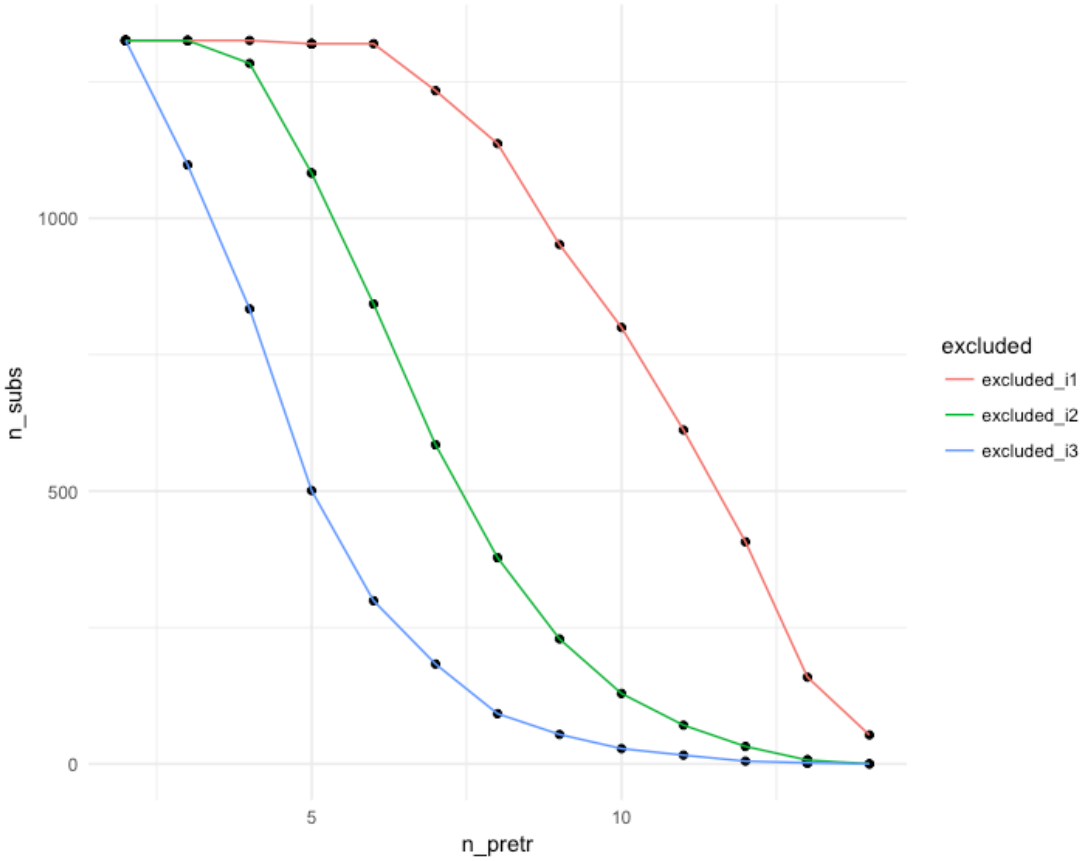


Figure: The x-axis is the number of pretreatment variables used for matching. The y-axis is the largest attainable sample size for a comparison of excluded versus not-excluded. "Excluded" was defined in three ways (see below). Curves i1-i3 reflect these increasingly stringent definitions of exclusion. These three definitions are based on different dichotomizations of the raw data depicted in the previous figure: ***excluded_i1*** = excluded by at least one person in wave 3, ***excluded_i2*** = same as before, but additionally not included by a single person in wave 3, ***excluded_i3*** = same as before but additionally not included by a single person in wave 1 either.

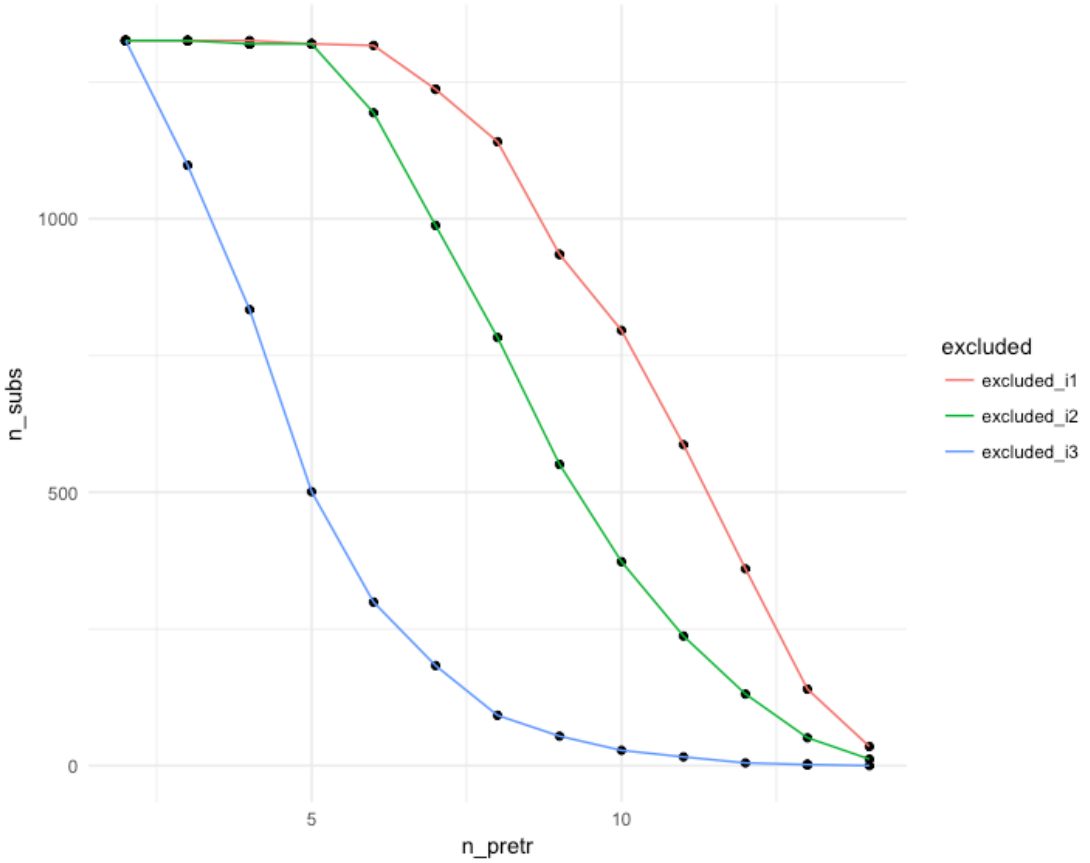


Figure: Same as previous but with exclusion/inclusion dichotomizations defined relative to the median exclusion/inclusion (not zero, as previously). This increases the sample size in every case, at the expense of a weaker treatment definition.