Bayesian Specification Curve Analysis

Supplementary Information

Christoph Semken David Rossell

27 October 2020

Contents

This document contains supplementary information for our paper "Bayesian Specification Curve Analysis". In Section 1 we briefly review the fundamentals behind Bayesian regression, Bayesian model selection (BMS) and Bayesian model averaging (BMA). In Section 2 we compare the properties of the BMA estimator and compare it to those of the median taken over all specifications. The simulation study also shows the needed R code to obtain inference on individual treatment/outcome combinations and on averaged treatment effects. In Section 3 we describe how BMS and BMA are used to produce a Bayesian Specification Curve Analysis (BSCA) and how to replicate our main results on teenager well-being using R. Section 4 contains supplementary data analyses that assess the robustness of our main findings for the Youth Risk Behavior Survey (YBRS) and Millenium Cohort Study (MCS) datasets. These robustness analyses include comparing linear to non-linear treatment effects and considering other well-being outcomes than those presented in the main paper. Section 5 describes certain aspects in which our analysis differs from that of Orben and Przybylski (2019). One motivation for these differences was to consider the inclusion of further control covariates, which as discussed in the main paper, is necessary to reduce biases in parameter estimates. We also adjusted the definition of some of the variables – e.g. expressing them on a common scale to facilitate interpretation and comparability, using validated psychometric scales as opposed to (unvalidated) individual outcome variables, fixing minor errors related to variable codes and including unemployed parents in the analysis.

1 Introduction to the Bayesian framework

This section provides a short introduction to the Bayesian regression, BMA and BMA frameworks. It is designed to allow readers who are unfamiliar with Bayesian statistics to follow the main text. We recommend anyone who wants to use BSCA to first get a deeper understanding of the Bayesian framework. Two excellent introductions are 'Statistical Rethinking' by McElreath (2020) and 'Bayesian Data Analysis' by Gelman et al. (2013). Two more applied texts are Hoeting et al.'s (1999) BMA tutorial and the open textbook (and course) 'An Introduction to Bayesian Thinking' by Clyde et al. (2020).

1.1 Bayesian regression

Bayesian statistics is based on a law of probability, known as Bayes theorem (or Bayes rule). It states that for two events A and B with non-zero probability, the probability of A occurring given that B occurred is

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)}$$

where $p(A)$ and $p(B)$ are the marginal (or unconditional) probabilities of A and B occurring, respectively.

To apply Bayes' rule to regression models, consider the standard linear regression model $y_i = \beta^T x_i + \alpha^T z_i + \epsilon_i$, where y_i is the value of the dependent variable, x_i contains one or more treatment variables of interest, z_i are adjustment covariates, and $\epsilon_i \sim N(0, \sigma^2)$ are independent errors for observations $i = 1, \dots, n$. The parameters $(\beta, \alpha, \sigma^2)$ describe

the probabilistic dependence of the outcome y_i on the treatment(s) x_i, after accounting for the effect of control covariates z_i. Specifically, β is a vector with one or more parameters that captures the association between the outcome and the treatment(s) of interest. α is a parameter vector capturing the association with control covariates which, despite not being of immediate interest, is needed to reduce biases and variance when estimating β. Although we describe the linear regression setting for simplicity, an analogous construction applies to any other regression model, including the logistic regression model used in the main paper.

Applying Bayes rule, the information about $(\beta, \alpha, \sigma^2)$ after observing the data is contained in a posterior probability distribution that is described by the density function

$$\overbrace{p(\beta, \alpha, \sigma^2 \mid \text{data})}^{\text{posterior prob.}} = \frac{\overbrace{p(\text{data} \mid \beta, \alpha, \sigma^2)}^{\text{likelihood}} \overbrace{p(\beta, \alpha, \sigma^2)}^{\text{prior prob.}}}{p(\text{data})}$$

where 'data' denotes the observed data $y_1, x_1, z_1, \ldots, y_n, x_n, z_n$. Values of $(\beta, \alpha, \sigma^2)$ receiving higher $p(\beta, \alpha, \sigma^2 \mid \text{data})$ are more supported, *a posteriori* after observing the data, than values receiving lower $p(\beta, \alpha, \sigma^2 \mid \text{data})$.

Two important quantities in the above equation are the likelihood function $p(\text{data} \mid \beta, \alpha, \sigma^2)$ and the prior distribution on the parameters $p(\beta, \alpha, \sigma^2)$. The likelihood is the probability (or probability density, for continuous data) that we would observe our actual data, given the parameters. The denominator $p(\text{data})$ does not depend on β, α, σ^2, it is a normalizing constant that we do not need to calculate directly and follows from the fact that (like all probability density functions) $p(\beta, \alpha, \sigma^2 \mid \text{data})$ integrates to 1.

The posterior information about our treatment effect coefficient(s) of interest β given the data is contained in $p(\beta \mid \text{data})$, which can be obtained by integrating the posterior distribution with respect to α and σ^2, that is

$$p(\beta \mid \text{data}) = \int p(\beta, \alpha, \sigma^2 \mid \text{data}) d\alpha d\sigma^2.$$

The posterior $p(\beta \mid \text{data})$ contains all the probabilistic information needed in a Bayesian analysis to make inference on the treatment effect(s) β. In particular, one may obtain a point estimate by taking the posterior mean of $p(\beta \mid \text{data})$, and quantify uncertainty via a 95% posterior interval (an interval that is assigned 95% probability by $p(\beta \mid \text{data})$). Of particular importance for BSCA, $p(\beta \mid \text{data})$ can be expressed via BMA as a weighted average across models, as we outline next.

1.2 Bayesian model selection

A common issue in regression analysis – and one that SCA and BSCA try to address – is that there are many potential treatment and control covariates a researcher can choose from. Specifically, for a total of $p = J+Q$ variables, where J is the number of treatments and Q the number of controls there are 2^p models $M_1, \ldots, M_{2^p}$, corresponding to the 2^p configurations of the variables in (x_i, z_i) that can potentially be included. Bayesian model selection (BMS) addresses this issue. It uses the posterior probability of the model to quantify the evidence in favor of a configuration of covariates being the optimal one (Kass and Raftery 1995). Here

optimal refers to dropping any variable such that its regression coefficient in α is truly equal to 0, while preserving variables that are truly associated with the outcome (i.e. having a non-zero entry in α). Two common strategies are then to either report estimates for the model with the highest posterior probability or use it to create weighted estimates. In Bayesian SCA we advocate for the latter, as described in the next section, so that any reported parameter estimates and intervals formally acknowledge the uncertainty in what is the right set of control variables.

The posterior probability of each model is given by Bayes' rule as

$$p(M_m \mid \text{data}) = \frac{p(\text{data} \mid M_m)p(M_m)}{p(\text{data})},$$

where $p(M_m)$ is a user-specified prior model probability and $p(\text{data} \mid M_m)$ is the so-called integrated likelihood (or marginal likelihood, or evidence) for model M_m. It can be computed using standard methods, either via closed-form expressions (when available), via deterministic approximations given by the Bayesian information criterion (BIC; Schwarz 1978), the extended BIC (EBIC; Chen and Chen 2008) or Laplace's method (Kass, Tierney, and Kadane 1990), or via stochastic approximations based on Markov Chain Monte Carlo methods (Friel and Wyse 2012).

For BSCA, again to avoid contentious prior choices, we use a uniform prior on the model size (the Beta-Binomial(1,1) prior; Scott and Berger 2010). This prior is uninformative in terms of how many covariates should be included in the regression. This also makes the estimation of the marginal likelihood computationally easy, since with these priors $p(\text{data} \mid M_m)p(M_m) \approx e^{-\frac{1}{2}\text{EBIC}_m}$, and hence we may use approximate posterior probabilities

$$p(M_m|\text{data}) \approx \frac{e^{-\frac{1}{2}\text{EBIC}_m}}{\sum_{m'=1}^{2^p} e^{-\frac{1}{2}\text{EBIC}_{m'}}}$$

where EBIC is the expected Bayesian information criterion (Chen and Chen 2008). Under fairly general conditions, the approximation becomes accurate as the sample size n grows (Schwarz 1978; Rossell 2018 Section 3). As a result, the BSCA model score is just a scaled log-transformation of the EBIC, which is a correction of the widely-used BIC to prevent the inclusion of false positives when the number of treatments J or controls Q are large.

1.3 Bayesian model averaging

In the Bayesian SCA, we use Bayesian model averaging (BMA) to get a combined estimate for each coefficient of interest from a large number of models, as well as 95% intervals that consider all possible specifications. Specifically, BMA expresses the posterior distribution as the weighted average

$$p(\beta, \alpha, \sigma^2 \mid \text{data}) = \sum_{m=1}^{2^p} p(\beta, \alpha, \sigma^2 \mid M_m, \text{data})p(M_m \mid \text{data}),$$

where $p(\beta, \alpha, \sigma^2 \mid M_m, \text{data})$ is the posterior distribution on $(\beta, \alpha, \sigma^2)$ after setting a subset of elements in α to zero, and $p(M_m \mid \text{data})$ is the posterior probability of the model M_m.

When the number of models 2^p is very large it is unfeasible to conduct the summation above exactly. In such situations one can use Markov Chain Monte Carlo methods, such as Gibbs sampling algorithms implemented in the R packages used in our examples.[1]

The point estimates and 95% intervals for treatment effects β – obtained from the BMA-weighted posterior $p(\beta \mid \text{data})$ – have several desirable properties. Under general conditions $p(M_m \mid \text{data})$ converges to 1 (as n grows) for the optimal model that only selects the covariates that are truly associated with the outcome (Dawid 1999). Under suitable mathematical conditions, this property holds even with large p and or if the data are not exactly generated by the assumed regression model – e.g. due to omitted covariates, non-linear effects, or other potentially incorrect parametric assumptions (Rossell 2018). As a result, the BMA point estimate and 95% interval converge to the frequentist MLE and 95% confidence interval obtained under the optimal model (LeCam 1953; Vaart 1998 Chapter 10, Theorem 10.1).[2]

1.4 Inference in BSCA

BSCA uses Bayesian regression and BMA to provide inference (point estimate, 95% posterior probability interval, and a hypothesis test) for each individual treatment-outcome combination, so that one can fully report their heterogeneity. BSCA can also provide point estimates and hypothesis tests on average treatment effects (ATE). In this section we outline how such inference is obtained.

We first discuss the BSCA for an individual treatment/outcome combination. Let β_j be the parameter quantifying said association. Inference is based on the posterior distribution $p(\beta_j \mid \text{data})$, shown in the BSCA top left panel. The panel also shows the posterior mean $E(\beta_j \mid \text{data})$ as a point estimate and a 95% posterior interval. This posterior distribution is also used to perform a hypothesis test for $\beta_j = 0$, specifically when the posterior probability $P(\beta_j \neq 0 \mid \text{data})$ exceeds a threshold one rejects $\beta_j = 0$. We recommend the threshold $P(\beta_j \neq 0 \mid \text{data}) > 0.95$ based on the property that, when the expected value of $P(\beta_j \neq 0 \mid \text{data})$ is above 0.95 (across repeated sampling), then the type I error is guaranteed to be below 0.05 (Rossell 2018 Corollary 1). Moreover, our EBIC-based formulation guarantees that, if truly an effect $\beta_j = 0$ is not present, then the expectation of $P(\beta_j \neq 0 \mid \text{data})$ and the type I error rate converge to 0, as n grows. See Section 1.5 for further discussion on false positive control.

The BSCA top right panel gives point estimates and 95% intervals given by the posterior distribution $p(\beta \mid M_m, \text{data})$ for each considered model (or the top 100 models, when the number of models 2^p is too large). This panel is analogous to standard SCA, except that we focus attention on the models more supported by the data. The BSCA middle panel gives the model scores, that is the posterior model probabilities $p(M_m \mid \text{data})$, for each configuration of adjustment covariates. Models are sorted decreasingly in terms of their posterior probabilities. The bottom panel mimicks that in SCA and indicates the covariates

[1] See also Madigan and Raftery (1994) for a description of BMS and Hoeting et al. (1999) for a tutorial on BMA.

[2] Put differently, it is possible to establish the mathematical validity (from a frequentist statistics viewpoint) of the posterior probabilities $p(M_m \mid \text{data})$. Briefly, this follows from the fact that $p(M_m \mid \text{data})$ converges to 1 for the optimal model and to 0 for any other model as n grows, Slutsky's theorem and standard Bernstein-von-Mises theorems.

that correspond to each model.

The BMA estimate and 95% interval take into account the uncertainty arising from the many possible model specifications, and is asymptotically valid from a frequentist point of view, as explained above. We emphasize that a main motivation for the original SCA was that standard errors conditional on a single selected model fail to account for the model selection uncertainty (Simonsohn, Simmons, and Nelson 2020). This issue is naturally resolved in BSCA by using the BMA weights. We neither have to pick one model (possible resulting in outcome reporting bias) nor take a simple average over all models (many of which maybe relatively implausible given the data, leading to potentially large estimation biases).

In addition, researchers might want to calculate an average treatment effect (ATE). We discuss first reporting an ATE across multiple treatments $\beta_1, \dots, \beta_J$, for a single outcome. It is common to define the ATE as the mean

$$\text{ATE} = \frac{1}{J}\sum_{j=1}^{J}\beta_j,$$

although it is also possible to consider other summaries, such as the median. Given a posterior distribution on the full β vector, $p(\beta \mid \text{data})$, one also has an implied posterior distribution $p(\text{ATE} \mid \text{data})$. It is hence straightforward to obtain a point estimate for the ATE via the posterior mean $E(\text{ATE} \mid \text{data})$, intervals with 95% posterior probability under $p(\text{ATE} \mid \text{data})$, and to reject the null hypothesis that ATE=0 when $P(\text{ATE} \neq 0 \mid \text{data}) > 0.95$. See the next section on how to obtain such inference in R. As described in the main paper, we recommend always reporting the individual treatment effects – at least their signs – when using the ATE.

Suppose now that one wishes to report a global ATE across $L > 1$ outcomes and $J \geq 1$ treatments. Let β_{jl} be the regression coefficient associated to treatment $j \in \{1, \dots, J\}$ and outcome $l = \{1, \dots, L\}$. SCA defines the ATE as the median β_{jl}, here we consider the (perhaps more standard) definition based on the mean

$$\text{ATE} = \frac{1}{JL}\sum_{l=1}^{L}\sum_{j=1}^{J}\beta_{jl}.$$

A point estimate for the global ATE is given by the posterior mean

$$E(\text{ATE} \mid \text{data}) = \frac{1}{JL}\sum_{l=1}^{L}\sum_{j=1}^{J}E(\beta_{jl} \mid \text{data}).$$

Note that for linear regression the global ATE associated to regressing each individual outcome on treatment and controls is mathematically equivalent to the ATE associated to regressing the average outcome on the treatments and controls. Section 2 illustrates how to exploit this property.

An important remark, which makes us caution against using the global ATE for statistical inference, is that it assigns equal weight to all outcomes. This may be inappropriate in situations where some of the outcomes are strongly correlated. For instance, suppose that there are $J = 10$ outcomes, 9 of which measure a very similar latent quantity (they are

similar items within a questionnaire) whereas the tenth outcome measures an inherently different quantity. The global ATE will be mostly determined by outcomes 1-9, whereas intuitively one might want to discount their weight. Given that defining alternative global ATE's is a potentially contentious issue, in our examples we use the standard ATE definition, and recommend that BSCA users rely on outcome-specific results.

1.5 Statistical considerations and false positive control

As discussed, our EBIC-based formulation guarantees that, as the sample size n grows, the total number of false discoveries across all tested treatment-outcome combinations converges to zero. See the simulation study in the next section for an empirical assessment of this property. In this section we discuss alternative Bayesian and non-Bayesian strategies to control false positives such as P-value adjustment and the False Discovery Rate.

Regarding alternative strategies, one could consider other Bayesian formulations that set different priors and still attain good properties (e.g. model selection consistency) as the sample size n grows. For instance, one could set so-called Complexity priors on the model space (Castillo, Schmidt-Hieber, and Vaart 2015) or non-local priors on the regression coefficients (Johnson and Rossell 2010, 2012; Rossell and Telesca 2017). These are recent developments in the statistical literature to further improve the prevention of false positives in settings with many parameters or hypothesis tests, but given the large sample size in our examples we found that these refinements were not needed. It is also possible to refine our formulation for situations where one considers many outcomes. Briefly, the discussed Beta-Binomial/EBIC property that the probability of having any false positive (family-wise error rate, FWER) converges to 0 for large n, for each individual outcome, implies that the overall FWER across outcomes also converges to 0. In our experience the Beta-Binomial/EBIC in practice attains a very good FWER control, as long as the number of outcomes is moderate relative to n. In situations with a truly large number L of outcomes one can extend the Beta-Binomial prior, by setting uniform prior probabilities to models that select $0, 1, \ldots, (J + q)L$ variables across all outcomes, where q is the number of potential control variables. While these refinements are potentially interesting, we found them to be unnecessary in the considered applications.

We remark that one can also use non-Bayesian false positive control methods. First, note that BMA could be viewed simply as a mechanism to obtain a point estimate, both a global $\widehat{\text{ATE}}$ and $\hat{\beta}_{jl} = E(\beta_{jl} \mid \text{data})$ for individual parameters. One can then use a permutation test akin to that used in SCA to obtain a P-value for the ATE. Permutation tests can also provide P-values for individual parameters, and one could use standard P-value adjustment or False Discovery Rate control methods to prevent false positive inflation due to testing multiple β_j's. See Benjamini and Hochberg (1995) and Efron (2007) for discussion on FWER and FDR control. Briefly, these methods ensure that the probability or proportion of false positives is below a pre-specified threshold, the default being the usual 0.05. That is, these methods are willing to admit a fixed positive probability of including regression parameters that are truly zero (similar to a P-value for a single test admitting a 0.05, say, false positive probability). In contrast the Beta-Binomial/EBIC formulation ensures that said probability converges to 0, that is as n grows one discards all truly irrelevant parameters, which intuitively provides a stronger false positive control.

2 BSCA estimator and hypothesis testing properties: a simulation study

We illustrate the use of BMA via a simple simulation study to infer individual treatment effects, as well as the average treatment effect (ATE). We evaluate the bias, root mean squared estimation error (RMSE) associated to the BMA point estimator, and the type I error probability (false positive) and power for testing if an effect is indeed present. We provide the R code so that readers can easily modify the simulation parameters.

We first summarize the results of the results for one outcome. In our settings BMA had near-zero bias and its RMSE was an order of magnitude smaller than SCA estimates. Further, the estimated type I error probability and power for all coefficients were near 0 and 1, respectively. These high-quality results are due to using a relatively large sample size of $n = 1000$. We chose this value because the teenager well-being studies also had such large n (actually larger), but we encourage readers to also assess performance in other settings. All scenarios include one control covariate that truly has an effect and is correlated with the treatment(s). In the multiple treatment case we consider effects of different magnitudes, to illustrate the power to detect smaller versus larger effects.

- Scenario 1 ($\beta_1 = 0$): `BMA bias = -7e-04, RMSE = 0.01, type I error = 0, SCA bias = 0.4979, RMSE = 0.4992`
- Scenario 2 ($\beta_1 = 1$): `BMA bias = -0.0027, RMSE = 0.044, power = 1, SCA bias = 0.4969, RMSE = 0.4984`
- Scenario 3 ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$): `BMA bias = 3e-04, RMSE = 0.0023, type I error = 0, SCA bias = 0.0751, RMSE = 0.0775`
- Scenario 4 ($\beta_1 = \beta_2 = 0, \beta_3 = 0.25, \beta_4 = 0.75, \beta_5 = \beta_6 = 1$): `BMA bias = -0.0027, RMSE = 0.0124, power = 1, SCA bias = 0.0549, RMSE = 0.0717`

Subsection 1 sets up some useful functions to run the simulations. Subsection 2 considers a simplest setting with a single treatment variable of interest. BMA and the SCA median are used to average inference across models, that is across possible specifications defined by each treatment/control having/not having an effect. Subsection 3 considers multiple treatments of interest. We show how to obtain inference for all individual treatment effects, so that one can assess and report on their heterogeneity. We also show how to estimate and test for the presence on an average treatment effect. Finally, Subsection 4 studies the properties of the BMA estimator in the case the ATE is taken over multiple outcomes.

2.1 Setup

We use the packages `mombf` and `specr` for BMA and SCA inference, respectively.

```
library(mombf)
library(specr)
```

We also load our auxiliary functions.

```
source('code/functions.R')
```

We create a function that simulates data from a Gaussian linear regression with n observations: the outcome variable y, treatment(s) x and control covariates z. The true value of the regression parameters for x and z are specified in β and α, respectively. The treatment(s) and controls are correlated, i.e. this is a situation where it is necessary to identify the relevant controls to avoid parameter estimation bias and increase the power of statistical tests.

```
simdata= function(n, beta, alpha, seed) {
 set.seed(seed)
 J= length(beta); q= length(alpha)
 x = matrix(rnorm(n*J),nrow=n) #simulate the value of treatment(s)
 xnames= paste('x',1:length(beta),sep='')
 if (q>0) { #if there are control covariates
   z = rowMeans(x) + matrix(rnorm(n*q),nrow=n) #correlated with treatments
   y = x %*% matrix(beta,ncol=1) + z %*% matrix(alpha,ncol=1) + rnorm(n)
   znames= paste('z',1:length(alpha),sep='')
   ans= data.frame(y,1,x,z)
   colnames(ans)= c('y','Intercept',xnames,znames)
 } else {   #if there are no control covariates
   y = x %*% matrix(beta,ncol=1) + rnorm(n)
   ans= data.frame(y,1,x)
   colnames(ans)= c('y','Intercept',xnames)
 }
 return(ans)
}
```

2.2 Single treatment

We consider two scenarios, both with a single treatment. In Scenario 1 the treatment has a truly zero effect, and a single control with a non-zero effect. This scenario assesses the type I error, that is the frequentist probability that BMA would wrongly claim the treatment effect to exist. In Scenario 2 the treatment has a truly non-zero effect, and we assess the statistical power of BMA to detect said effect.

2.2.1 Truly no treatment effect

BMA point estimates are stored in `bmaest` and its posterior probabilities on the presence of an effect in `margpp`. We also store in `scaest` the point estimate returned by a standard Specification Curve Analysis, which takes the median across all possible specifications.

```
n= 1000; beta= 0; alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
```

```
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=300+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1],verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b= coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")
    scaest[i]= summarise_specs(unique(sca))$median
}
```

We report the estimated bias and RMSE. The BMA estimate has a bias close to zero, in contrast SCA tends to over-estimate the true parameter value $\beta_1 = 0$. The reason is that x is correlated with the control z, which truly has an effect, hence the model including x but not z over-estimates β_1 (this follows from simple algebra and standard least-squares theory).

We also assess the type I error for testing the null hypothesis $\beta_1 = 0$ versus the alternative $\beta_1 \neq 0$. The BMA test rejects $\beta_1 = 0$ when the posterior probability $P(\beta_1 \neq 0 \mid y)$ is large, specifically $P(\beta \neq 0 \mid y) > 0.95$ guarantees that the type I error is below 0.05 (as the sample size $n \to \infty$). In most simulations $P(\beta \neq 0 \mid y)$ took a pretty small value (run `summary(margpp[,'x1'])` in R) and the null hypothesis was never rejected, i.e. the estimated type I error is 0.

```
bma.bias= mean(bmaest[,'x1'] - beta)
bma.rmse= sqrt(mean((bmaest[,'x1'] - beta)^2))
bma.reject= mean(margpp[,'x1']>0.95)
sca.bias= mean(scaest - beta)
sca.rmse= sqrt(mean((scaest - beta)^2))
tab1= rbind(c(bma.bias, bma.rmse, bma.reject), c(sca.bias, sca.rmse, NA))
rownames(tab1)= c('BMA','SCA')
colnames(tab1)= c('Bias','RMSE','type I error')
tab1
```

	Bias	RMSE	type I error
BMA	-0.001	0.010	0
SCA	0.498	0.499	NA

2.2.2 Truly non-zero treatment effect

We repeat the exercise with a non-zero treatment effect $\beta_1 = 1$. The table below repors the estimated bias, RMSE and power to detect that truly $\beta_1 \neq 0$. The null hypothesis is rejected in all simulations, hence the estimated power is 1.

```
n= 1000; beta= 1; alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=100+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b= coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")[-1,]
    scaest[i]= summarise_specs(sca)$median
}
```

```
bma.bias= mean(bmaest[,'x1'] - beta)
bma.rmse= sqrt(mean((bmaest[,'x1'] - beta)^2))
bma.reject= mean(margpp[,'x1']>0.95)
sca.bias= mean(scaest - beta)
sca.rmse= sqrt(mean((scaest - beta)^2))
tab2= rbind(c(bma.bias, bma.rmse, bma.reject), c(sca.bias, sca.rmse, NA))
rownames(tab2)= c('BMA','SCA'); colnames(tab2)= c('Bias','RMSE','Power')
tab2
```

	Bias	RMSE	Power
BMA	-0.003	0.044	1
SCA	0.497	0.498	NA

2.3 Multiple treatments

An interesting feature of SCA is visualizing the heterogeneity across multiple treatments/outcomes, and providing an averaged (or median) treatment effect over treat-

ments/outcomes (and sets of control covariates). We show how to use BMA to estimate and test individual treatments, so that one can distinguish those with positive/zero/negative effects, as well as estimating an testing for the presence of an average treatment effect.

For simplicity we show an example with a single outcome and 6 treatments. We first consider a setting where all 6 treatments truly have a zero effect, so the ATE=0. This setting helps assess the probability that any individual treatment is falsely declared to have an effect. We then consider a second setting where 3 treatments truly have an effect and the remaining 2 treatments do not, which helps assess the statistical power of BMA tests on individual treatments, and on the ATE. More specifically, the ATE is usually defined as ATE=$(\sum_{j=1}^{6} \beta_j)/6$. We test this null hypothesis by obtaining the posterior probability $P(\text{ATE} \neq 0 \mid y) = P(\sum_j \beta_j \neq 0 \mid y)$, and rejecting the null hypothesis when said probability exceeds the 0.95 threshold.

We remark on an important property of our specific BMS implementation. In classical P-value based hypothesis tests the probability of claiming at least one false positive finding (family-wise type I error) increases as one adds more treatments, unless one uses more stringent P-value thresholds. In our BMS framework false positive inflation is avoided by using a Beta-Binomial(1,1) prior on the model space (or the EBIC approximation to model posterior probabilities, as outlined earlier). This formulation adds a penalization term as one increases the number of treatments (or covariates). The penalization ensures that as the sample size n grows, the family-wise type I error probability converges to 0, see Chen and Chen (2008) or Rossell (2018) (Sections 3-4) for a mathematical proof. It also ensures that BMA parameter estimates and 95% intervals converge to those obtained via maximum likelihood estimation under the model that includes only the subset of treatments and controls truly associated with the outcome Rossell and Telesca (2017) (Proposition 3).

2.3.1 Zero average treatment effect

The simulation is as before, with the difference that `beta` has now 6 elements.

```
n= 1000; beta= c(0,0,0,0,0,0); alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=200+i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b = coef(ms)
```

```
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #ATE
    bma.ate[i,]= getATE(ms, xvars=xnames, fun='mean')[c('estimate','margpp')]
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")
    scaest[i]= summarise_specs(unique(sca))$median
}
```

We first report the bias, RMSE and type I error probabilities for individual treatments. BMA did not declare as significant any individual treatment in any of the simulated datasets, that is both the individual and family-wise type I error probabilities are estimated to be near-zero. The individual treatment effects are not usually considered in SCA analyses.

```
bias.indiv= rowMeans(t(bmaest[,xnames]) - beta)
rmse.indiv= sqrt(rowMeans((t(bmaest[,xnames]) - beta)^2))
reject.indiv= colSums(margpp[,xnames] > 0.95) / nrow(margpp)
tab3.indiv= rbind(bias.indiv, rmse.indiv, reject.indiv)
rownames(tab3.indiv)= c('Bias','RMSE','type I error')
round(tab3.indiv, 5)
```

	x1	x2	x3	x4	x5	x6
Bias	0.000	0.000	0.001	0.000	0.000	0.001
RMSE	0.003	0.006	0.009	0.003	0.003	0.009
type I error	0.000	0.000	0.000	0.000	0.000	0.000

We next estimate the bias, RMSE and type I error associated to the ATE. In this simulation, the absolute bias and RMSE are more than 100 and 10 times larger for the median taken over all specifications, respectively.

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)

sca.biasate= mean(scaest - ate)
sca.rmseate= sqrt(mean((scaest - ate)^2))
tab3= rbind(c(bma.biasate, bma.rmseate, bma.rejectate),
            c(sca.biasate, sca.rmseate, NA))
rownames(tab3)= c('BMA','SCA')
colnames(tab3)= c('Bias','RMSE','type I error')
round(tab3, 4)
```

	Bias	RMSE	type I error
BMA	0.000	0.002	0
SCA	0.075	0.078	NA

2.3.2 Non-zero average treatment effect

Lastly we consider a setting where 2 treatments truly have no effect ($\beta_1 = \beta_2 = 0$) and 4 treatments have heterogeneous effects ($\beta_3 = 0.25$, $\beta_4 = 0.75$, $\beta_4 = 1$, $\beta_5 = 1$). Again, we include one control covariate that is correlated with the treatments and truly has an effect. Note that the true ATE and median treatment effects are both 0.5, to facilitate comparison between the BMA and SCA results. This is to facilitate comparison between BMA and SCA , since in our implementation BMA targets the ATE and SCA the median. One can use BMA to obtain inference on the median by using `getATE(ms, xvars=xnames, fun='median')` below.

```
n= 1000; beta= c(0,0,1/4,3/4,1,1); alpha= 1; nsims= 100
xnames= paste('x',1:length(beta),sep='')
znames= paste('z',1:length(alpha),sep='')
bmaest= margpp= matrix(NA,nrow=nsims,ncol=1+length(beta)+length(alpha))
colnames(bmaest)= colnames(margpp)= c('Intercept', xnames, znames)
scaest= double(nsims)
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
for (i in 1:nsims) {
    data= simdata(n=n,beta=beta,alpha=alpha,seed=i)
    #BMA
    ms = modelSelection(
      y=data[,1], x=data[,-1], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    b = coef(ms)
    bmaest[i,]= b[-nrow(b),'estimate']
    margpp[i,]= b[-nrow(b),'margpp']
    #ATE
    bma.ate[i,]= getATE(ms, xvars=xnames, fun='mean')[c('estimate','margpp')]
    #SCA
    sca= run_specs(df=data, y="y", x=xnames, controls=znames, model="lm")
    scaest[i]= summarise_specs(unique(sca))$median
}
```

We first report the bias, RMSE and type I error probabilities for individual treatments. BMA correctly detected that $\beta_1 = \beta_2 = 0$, and that $\beta_3 \neq 0$, $\beta_4 \neq 0$, $\beta_5 \neq 0$ and $\beta_6 \neq 0$ in all simulations.

```
bias.indiv= rowMeans(t(bmaest[,xnames]) - beta)
rmse.indiv= sqrt(rowMeans((t(bmaest[,xnames]) - beta)^2))
reject.indiv= colSums(margpp[,xnames] > 0.95) / nrow(margpp)
tab4.indiv= rbind(bias.indiv, rmse.indiv, reject.indiv)
rownames(tab4.indiv)= c('Bias','RMSE','Proportion rejected')
round(tab4.indiv, 5)
```

	x1	x2	x3	x4	x5	x6
Bias	-0.001	-0.002	-0.002	-0.006	-0.004	-0.001
RMSE	0.010	0.016	0.030	0.034	0.033	0.035
Proportion rejected	0.000	0.010	1.000	1.000	1.000	1.000

We next estimate the bias, RMSE and type I error associated to the ATE. Here BMA correctly detected that the ATE$\neq 0$ in all simulations.
Again, the bias and RMSE are magnitudes larger for the median taken over all specifications than for the BMA average.

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)

sca.biasate= mean(scaest - ate)
sca.rmseate= sqrt(mean((scaest - ate)^2))
tab4= rbind(c(bma.biasate, bma.rmseate, bma.rejectate),
            c(sca.biasate, sca.rmseate, NA))
rownames(tab4)= c('BMA','SCA')
colnames(tab4)= c('Bias','RMSE','Power')
round(tab4, 4)
```

	Bias	RMSE	Power
BMA	-0.003	0.012	1
SCA	0.055	0.072	NA

2.4 Multiple outcomes

Although we generally recommend running BSCA for each outcome individually and reporting the whole heterogeneity across outcomes, below we illustrate how to perform inference for a global ATE across L outcomes and J treatments. We define the global ATE in terms of the original outcomes y,

$$\text{ATE} = \frac{1}{JL}\sum_{l=1}^{L}\sum_{j=1}^{J}\beta_{lj}$$

where β_{jl} is the regression parameter for treatment j on outcome l.

We consider a setting with $L = 4$ outcomes, $J = 5$ treatments and $q = 1$ control variable. The outcomes are generated from a multivariate linear regression model where the errors are correlated, specifically $\epsilon \sim N(0, \Sigma)$ where Σ has unit variances in the diagonal, pairwise correlations equal to 0.9 among outcomes 1-3, and 0.1 correlation with outcome 4. This correlation structure is meant to represent a situation where the first three outcomes can be thought of as measuring one common latent chararacteristic, that is different from that measured by the fourth outcome.

```
L= 4; J= 5; q= 1; n= 1000; nsims= 100
Sigma= diag(L)
Sigma[1,2]= Sigma[1,3]= Sigma[2,3]= Sigma[2,1]= Sigma[3,1]= Sigma[3,2]= 0.9
Sigma[1:3,4]= 0.1; Sigma[4,1:3]= 0.1
```

Function `simmultivdata` (created in the supplementary file `functions.R`) generates such data y.

2.4.1 No average treatment effect

Our first simulation considers a setting where no treatment truly has an effect, hence ATE = 0. The bias, RMSE to estimate ATE_y are reported below, as well as the type I error rate, which is essentially zero.

```
beta= matrix(0,nrow=J,ncol=L); alpha= matrix(c(1,1,1,-1),nrow=q, ncol=L)

bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
xnames= paste('x',1:J,sep='')
for (i in 1:nsims) {
    data= simmultivdata(n=n,beta=beta,alpha=alpha,Sigma=Sigma,seed=300+i)
    y= as.matrix(data[, 1:L])
    m= rowMeans(y)
    ms = modelSelection(
      y=m, x=data[,-1:-L], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    bma.ate[i,]= getATE(ms, xvars=xnames)[c('estimate','margpp')]
}
```

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)
```

```
tab5= c(bma.biasate, bma.rmseate, bma.rejectate)
names(tab5)= c('Bias','RMSE','Type I error')
round(tab5, 4)
```

```
##         Bias         RMSE Type I error
##       0.0000       0.0018       0.0000
```

2.4.2 Non-zero average treatment effect

Next we consider a case where 3 treatments truly have an effect, whereas treatments 4-5 do not. The effect of treatments 1-3 on outcomes 1-3 is different to that on outcome 4, again mimicking a situation where outcomes 1-3 measure a common latent characteristic.

```
beta= matrix(0,nrow=J,ncol=L)
beta[,1]= beta[,2]= beta[,3]= c(1,1,1,0,0)
beta[,4]= c(1/4, 1/4, 1/4, 0, 0)
alpha= matrix(c(1,1,1,-1),nrow=q, ncol=L)
```

```
bma.ate= matrix(NA, nrow=nsims, ncol=2)
colnames(bma.ate)= c('estimate','margpp')
xnames= paste('x',1:J,sep='')
for (i in 1:nsims) {
    data= simmultivdata(n=n,beta=beta,alpha=alpha,Sigma=Sigma,seed=300+i)
    y= as.matrix(data[, 1:L])
    m= rowMeans(y)
    ms = modelSelection(
      y=m, x=data[,-1:-L], verbose=FALSE,
      priorCoef=zellnerprior(tau=nrow(data))
    )
    bma.ate[i,]= getATE(ms, xvars=xnames)[c('estimate','margpp')]
}
```

```
ate= mean(beta)
bma.biasate= mean(bma.ate[,'estimate'] - ate)
bma.rmseate= sqrt(mean((bma.ate[,'estimate'] - ate)^2))
bma.rejectate= mean(bma.ate[,'margpp']>0.95)
```

```
tab6= c(bma.biasate, bma.rmseate, bma.rejectate)
names(tab6)= c('Bias','RMSE','Power')
round(tab6, 4)
```

```
##    Bias    RMSE   Power
## -0.0010  0.0089  1.0000
```

3 Reproducing the BSCA results for the teenager well-being datasets

We explain how to obtain BMA and BSCA results in R by reproducing our findings in the two teenager well-being datasets. Subsection 1 loads R packages and functions needed for the analysis, as well as the Youth Risk Behavior Survey (YBRS) and Millenium Cohort Study (MCS) datasets on adolescent mental well-being and technology use. Subsection 2 applies Bayesian model selection and averaging to both datasets. Finally, Subsections 3 and 4 produce Figs. 1 (single-outcome BSCA) and 2 (multiple-outcome BSCA) presented in the manuscript, respectively.

3.1 Setup

We start by loading the required R packages. For the statistical analysis we use `mombf` and `BAS`, whereas `tidyverse` offers some convenient functions for treating the data. We source file `functions.R`, which contains auxiliary functions to produce the BSCA plots, and load the pre-processed version of the YBRS and MCS datasets in files `yrbs.Rdata` and `mcs.Rdata`, respectively. We cannot provide these pre-processed data due to copyright issues, but you can run the code provided in the replication file `bsca_preanalysis.Rmd` to create the processed data from the raw data and code provided by Orben and Przybylski (2019). See also `bsca_preanalysis.html` for a compiled version displaying the R code and output.

```
library(mombf)
library(BAS)
library(tidyverse)
```

```
source('code/functions.R')
yrbs = new.env(); mcs = new.env()
load('data/export/yrbs.Rdata', yrbs)
load('data/export/mcs.Rdata', mcs)
```

3.2 Bayesian model selection

3.2.1 YRBS data

We begin by analyzing the YRBS data. The data frame `y` stores several outcomes, whereas `x` stores treatment variables and `data` stores other recorded variables. We specify that we wish to use the outcome variable *thought about suicide* (the second column in `y`) by setting `idy=2`. We also specify that we wish to use TV and electronic device use (first and second column in `x`) as treatment variables to be jointly included in all regression models by setting `idx=c(1,2)`. Finally we specify to use race, age, gender, school grade, survey year and body mass index as potential control variables (saved in `cvars` and `cvarsplus`).

```
attach(yrbs)
names(y)
```

```
## [1] "loneliness"     "think suicide"  "plan suicide"   "commit suicide"
## [5] "doctor suicide"
```

```
names(x)
```

```
## [1] "TV Use"                "Electronic Device Use"
```

```
c_names
```

```
## [1] "Race"    "Aged 12" "Aged 13" "Age"     "Sex"     "Grade"   "Year"
## [8] "BMI"
```

```
idy = 2; idx = c(1,2)
datareg = data.frame(y[,idy], x[,idx], data[,c(cvars,cvarsplus)])
names(datareg) = c('y', names(x)[idx], c_names)  # set names
datareg = datareg[rowSums(is.na(datareg))==0, ]  # remove NAs
detach(yrbs)
```

The data.frame `datareg` contains the outcome, treatment and control variables. For illustration, its first few rows are displayed below. These variables have been conveniently coded so they can be entered directly into the usual R regression equation. For instance, `Aged 12` and `Aged 13` are indicators for an individual's age being 12 and 13 years, whereas `Age` contains the age in years, using these 3 columns to code the effect of age allows to capture non-linear effects detected in preliminary exploratory data analyses (see reproduction file `bsca_preanalysis.Rmd`).

```
head(datareg %>% rename('ED Use' = 'Electronic Device Use'))
```

	y	TV Use	ED Use	Race	Aged 12	Aged 13	Age	Sex	Grade	Year	BMI
45	0	medium	0.333	0	1	0	12	1	11	2007	20.85
46	0	low	0.000	0	1	0	12	1	12	2007	23.85
47	0	medium	0.500	0	1	0	12	1	9	2007	18.04
50	1	high	1.000	0	1	0	12	1	12	2007	29.47
54	0	low	0.167	0	1	0	12	1	11	2007	26.36
55	0	high	0.000	0	0	1	13	1	9	2007	21.19

A first step in BSCA is to run Bayesian model selection (BMS), which will assign a score (posterior probability) to each model (possible set of control variables). Next, we

use Bayesian model averging (BMA) to combine these estimates. Since the outcome variable is binary we use logistic regression models, setting a uniform prior on the model size (`modelbbprior(1,1)`). The function `mombf:::modelSelectionGLM` computes scores for all 1024 possible models, and may take a while to run.

```
yrbs_ms = mombf:::modelSelectionGLM(
  y ~ ., data=datareg,
  includevars=1, familyglm=binomial(link='logit'),
  priorDelta=modelbbprior(1,1)
)
```

```
## Enumerating 1024 models.........
```

The BMA treatment effect estimates, 95% posterior intervals and marginal posterior probability that the variable has an effect on the outcome are stored in `yrbs_coef`. The output indicates a strong evidence that both treatments have an effect, albeit with opposing signs, and that age, gender, grade and BMI are control covariates that one should include in the model to avoid biases in the treatment effect estimates (driven by under-selection of truly relevant controls).

```
yrbs_coef = coef(yrbs_ms)
options(scipen=999)  # turn off scientific notation
yrbs_coef
```

	estimate	2.5%	97.5%	margpp
(Intercept)	-1.343	-1.757	-1.073	1.000
‘TV Use‘medium	-0.233	-0.281	-0.187	1.000
‘TV Use‘high	-0.047	-0.121	0.027	1.000
‘Electronic Device Use‘	0.629	0.566	0.692	1.000
Race	0.001	0.000	0.026	0.036
‘Aged 12‘	1.144	0.000	2.171	0.790
‘Aged 13‘	-0.002	0.000	0.000	0.009
Age	0.011	0.000	0.073	0.221
Sex	-0.789	-0.833	-0.744	1.000
Grade	-0.086	-0.148	-0.057	1.000
Year2009	0.000	0.000	0.000	0.006
Year2011	0.000	0.000	0.000	0.006
Year2013	0.000	0.000	0.000	0.006
Year2015	0.001	0.000	0.000	0.006
BMI	0.025	0.021	0.029	1.000

Given these coefficients, we use the function `getOR` to obtain odds ratios for increasing the TV use from low to medium/high, and for increasing the electronic device use from 0 to

>= 5 hours (coded as EDU=1 in our dataset, leading to setting `treatvals=1` below). The function exponentiates the coefficient estimates and formats the result.

```
getOR(yrbs_coef, treat='TV Use', digits=2)
```

	OR	CI.low	CI.up
'TV Use'medium	0.79	0.76	0.83
'TV Use'high	0.95	0.89	1.03

```
getOR(yrbs_coef, treat='Electronic Device Use', treatvals=1, digits=2)
```

	OR	CI.low	CI.up
Electronic Device Use 1	1.88	1.76	2

Note that the regression models include simultaneously the two treatment variables, TV and electronic device (ED) usage, which is necessary to avoid biased estimates when treatments are correlated. In these data the correlation is mild, for instance Pearson's correlation between the number of hours of TV use and ED use (columns `q81` and `q82` in `data`) is 0.21.

```
round(cor(na.omit(yrbs$data[c('q81', 'q82')]), method='pearson')[1,2],
      digits=2)
```

```
## [1] 0.21
```

3.2.2 MCS data

Next, we load the MCS dataset and run Bayesian model selection and averaging, analogously to the above analysis of the YRBS data. As described in the main manuscript, for illustration in our analysis we considered 4 outcome variables, 5 treatments and 14 potential control variables (results for other outcome variables are shown in Fig. S1). We display the names of these variables, which have been stored in the `mcs` workspace.

```
attach(mcs)
```

```
names(yvars)
```

```
## [1] "Depressed (adolescent)"
## [2] "Low self-esteem (adolescent)"
## [3] "High total difficulties (parent)"
## [4] "High emotional problems (parent)"
```

```
## [5] "High conduct problems (parent)"
## [6] "High hyperactivity/inattention (parent)"
## [7] "High peer problems (parent)"
## [8] "Low pro-sociality (parent)"
```

```
x_names
```

```
## [1] "TV"               "Electronic games" "Social media"     "Other internet"
## [5] "Own computer"
```

```
names(cvars)
```

```
##  [1] "Male"       "Age"        "BMI"        "Motivation" "Ethnicity"
##  [6] "Closeness"  "Father"     "Score"      "Employed"   "Illness"
## [11] "Time"       "Distress"   "Siblings"   "Income"
```

The code below runs BMA for all 5 outcome variables. One option is to use the `BAS` package, which implements MCMC sampling to explore the model space (sets of control covariates to be potentially included in the regression). A faster alternative analysis is to set `fast = TRUE`, which limits the analysis to the top 100 models (i.e. those shown in the BSCA plot) according to a prior screening (stored in `pp_lin`, see `bsca_preanalysis.Rmd` for details). The posterior probabilities of any model after the 100 first ones is vanishingly small, and the final BSCA results are virtually identical to those of the analysis using the `BAS` package.

```
fast = TRUE

mcs_coef = list(); mcs_ms = list()
for (idy in 1:length(yvars)) {
  # select data
  yvar = yvars[idy]; yname = names(yvars)[idy]
  datareg = na.omit(data[c(yvar, x_vars, cvars)])
  names(datareg) = c('y', x_names, names(cvars))

  # BMA
  if (fast) {
    models = as.matrix.pp(
      pp_lin[[idy]], nummodels=100, numvars=length(datareg)
    )
    mcs_ms[[yname]] = mombf:::modelSelectionGLM(
      y ~ ., data=datareg, models=models,
      familyglm= binomial(link='logit'),
      priorDelta=modelbbprior(1,1)
    ); cat('\n')
```

```
  } else {
    mcs_ms[[yname]] = BAS:::bas.glm(
      y~., data=datareg, family=binomial(link='logit'),
      betaprior=bic.prior(), modelprior=beta.binomial(),
      method='MCMC', n.models=150
    )
  }
  mcs_coef[[yname]] = coef(mcs_ms[[yname]])
}
```

```
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
## Enumerating 100 models.........
```

```
detach(mcs)
```

We can inspect the BMA results for the four outcomes. For brevity here we focus on *parent-assessed* high total difficulties and *adolescent-assessed* depression (see Subsection 3 below for a plot summarizing the BMA results for the two other outcomes). The analysis provides strong evidence that, according to parents, social media decrease the odds of total difficulties (marginal posterior probability=1, up to rounding), whereas electronic games increase those odds. We also find strong evidence that BMI, educational motivation, closeness to parents, the primary caregiver's word ability score and psychological distress, presence of a longstanding illness and the household income are necessary control covariates, as well as strong evidence that other covariates are not needed.

```
mcs_coef[["High total difficulties (parent)"]]
```

	estimate	2.5%	97.5%	margpp
(Intercept)	1.895	1.027	2.797	1.000
TV	-0.005	0.000	0.000	0.017
'Electronic games'	0.572	0.353	0.794	1.000
'Social media'	-0.719	-0.960	-0.477	1.000
'Other internet'	0.000	0.000	0.000	0.004
'Own computer'	0.000	0.000	0.000	0.003
Male	0.000	0.000	0.000	0.005
Age	-0.003	-0.002	0.000	0.028
BMI	0.043	0.027	0.058	1.000
Motivation	-0.789	-0.936	-0.638	1.000
Ethnicity	0.000	0.000	0.000	0.003
Closeness	-0.537	-0.650	-0.425	1.000
Father	0.048	0.000	0.303	0.244
Score	-0.055	-0.074	-0.036	1.000
Employed	-0.004	-0.078	0.000	0.032
Illness	1.037	0.877	1.197	1.000
Time	-0.012	-0.124	0.000	0.128
Distress	0.113	0.098	0.128	1.000
Siblings	-0.001	-0.014	0.000	0.031
Income	-0.001	-0.002	-0.001	1.000

The output for adolescent-assessed depression can be interpreted analogously. Briefly, here there is strong evidence that social media and other internet use increase the odds of depression, in contrast with the results of the earlier parent-assessed outcome. Some of the needed control covariates are also different, for instance males self-report lower odds of depression than females, whereas gender did not play a role in the parental assessment.

```
mcs_coef[["Depressed (adolescent)"]]
```

	estimate	2.5%	97.5%	margpp
(Intercept)	4.105	3.238	4.982	1.000
TV	0.000	0.000	0.000	0.005
'Electronic games'	0.006	0.000	0.056	0.028
'Social media'	1.001	0.647	1.349	1.000
'Other internet'	1.037	0.664	1.404	1.000
'Own computer'	-0.001	0.000	0.000	0.008
Male	-1.114	-1.268	-0.954	1.000
Age	0.001	0.000	0.000	0.014
BMI	0.045	0.028	0.061	1.000
Motivation	-1.607	-1.763	-1.455	1.000
Ethnicity	-0.002	0.000	0.000	0.018
Closeness	-0.783	-0.896	-0.672	1.000
Father	0.000	0.000	0.000	0.008
Score	0.013	0.000	0.041	0.501
Employed	0.000	0.000	0.000	0.005
Illness	0.511	0.337	0.689	1.000
Time	0.000	0.000	0.000	0.009
Distress	0.002	0.000	0.026	0.109
Siblings	-0.009	-0.106	0.000	0.110
Income	0.000	0.000	0.001	0.086

To summarize the treatment effects of interest we again use the auxiliary function `getOR`. These correspond to odds ratios for increasing the use of social media from 0 (no usage) to >7 hours (coded as 1 in our dataset, hence we set `treatvals=1`). We first obtain odds-ratios and 95% posterior intervals for social media and electronic games on parent-assessed total difficulties.

```
getOR(mcs_coef$`High total difficulties (parent)`, treat='Social media',
      treatvals=1, digits=2)
```

	OR	CI.low	CI.up
Social media 1	0.49	0.38	0.62

Next we report the odds ratios for adolescent self-assessed depression and low self-esteem.

```
getOR(mcs_coef$`Depressed (adolescent)`, treat='Social media',
      treatvals=1, digits=2)
```

	OR	CI.low	CI.up
Social media 1	2.72	1.91	3.85

```
getOR(mcs_coef$`Low self-esteem (adolescent)`, treat='Social media',
      treatvals=1, digits=2)
```

	OR	CI.low	CI.up
Social media 1	1.52	1	2.73

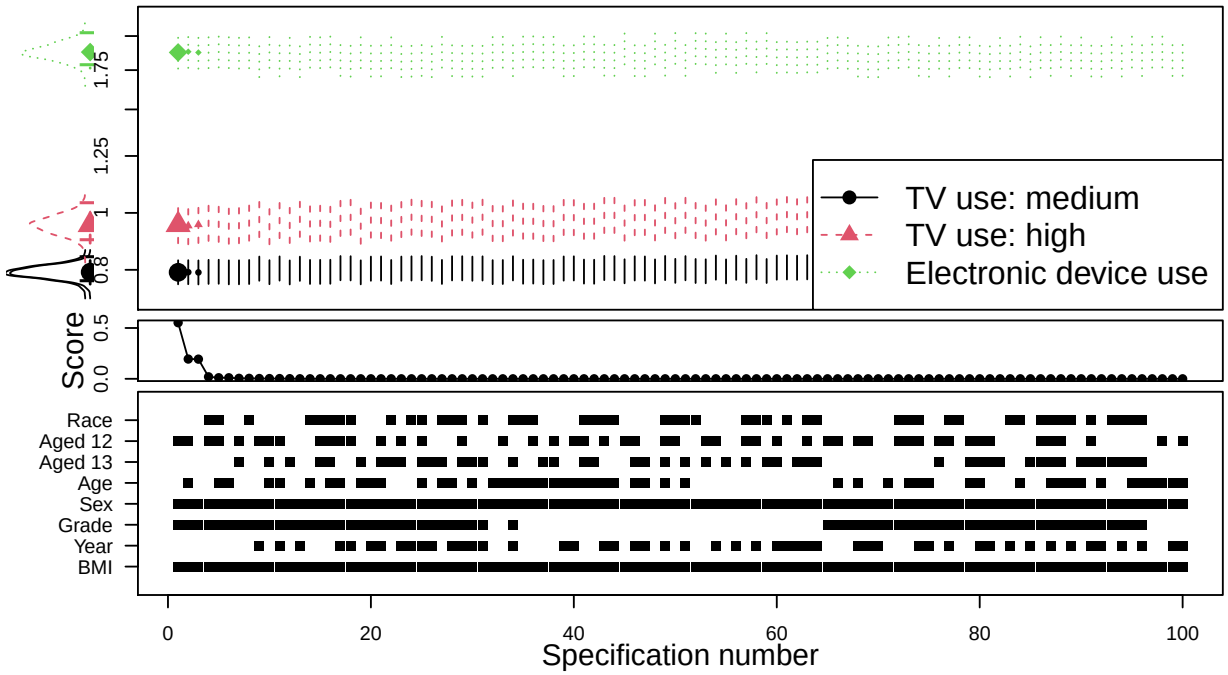
3.3 Reproducing Figure 1 (single-outcome BSCA)

3.3.1 Top panel: YRBS data

We use function `single_bsca` to plot the single-outcome BSCA for the YRBS data and the outcome *thought about suicide*. The argument `coefidx` specifies the names of the treatment variables that should be plotted. The function also allows specifying optional arguments such as the treatment names to be displayed in the legend (argument `x.labels`), variable names to be displayed in the bottom panel displaying variable configurations (argument `var.labels`), and omitting variables from that panel (argument `omitvars`, useful when there are many variables or when several columns code for the non-linear effect of a single variable and are always included together, such as year in the YRBS data). The labels on the y axis are stored in an array whose names (optionally) are the original values (argument `y.labels`). Here, we turn the estimated coefficient into the odds ratio by exponentiating it.

```
idx_fit = c(2:4)
id_years = c(12:14)
y_labels = c(0.8, 1, 1.25, 1.5, 1.75, 2)
names(y_labels) = log(y_labels)  # y scale as odds ratio

single_bsca(
  yrbs_ms, coefidx=idx_fit, omitvars=c(1, idx_fit, id_years),
  x.labels=yrbs$x_labels, var.labels=yrbs$c_names, y.labels=y_labels
)
```

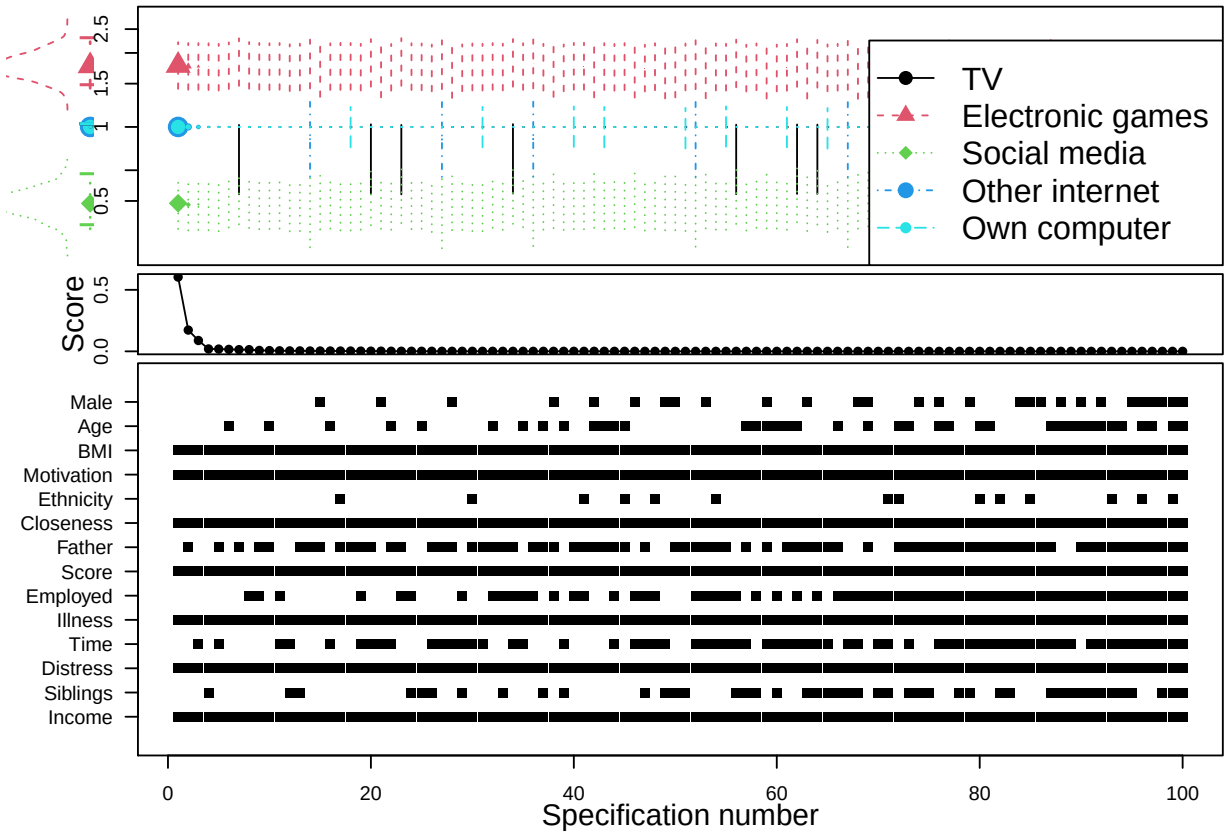


3.3.2 Bottom panel: MCS data

Similarly, we plot the single-outcome BSCA for the MCS data and the parent-assessed outcome *high total difficulties*.

```
y_labels = c(1/2, 1/1.5, 1, 1.5, 2, 2.5)
names(y_labels) = log(y_labels)  # y scale as odds ratio

single_bsca(
  mcs_ms$`High total difficulties (parent)`, coefidx=2:6,
  x.labels=mcs$x_names, var.labels=names(mcs$cvars), y.labels=y_labels,
  height.vars=0.55
)
```



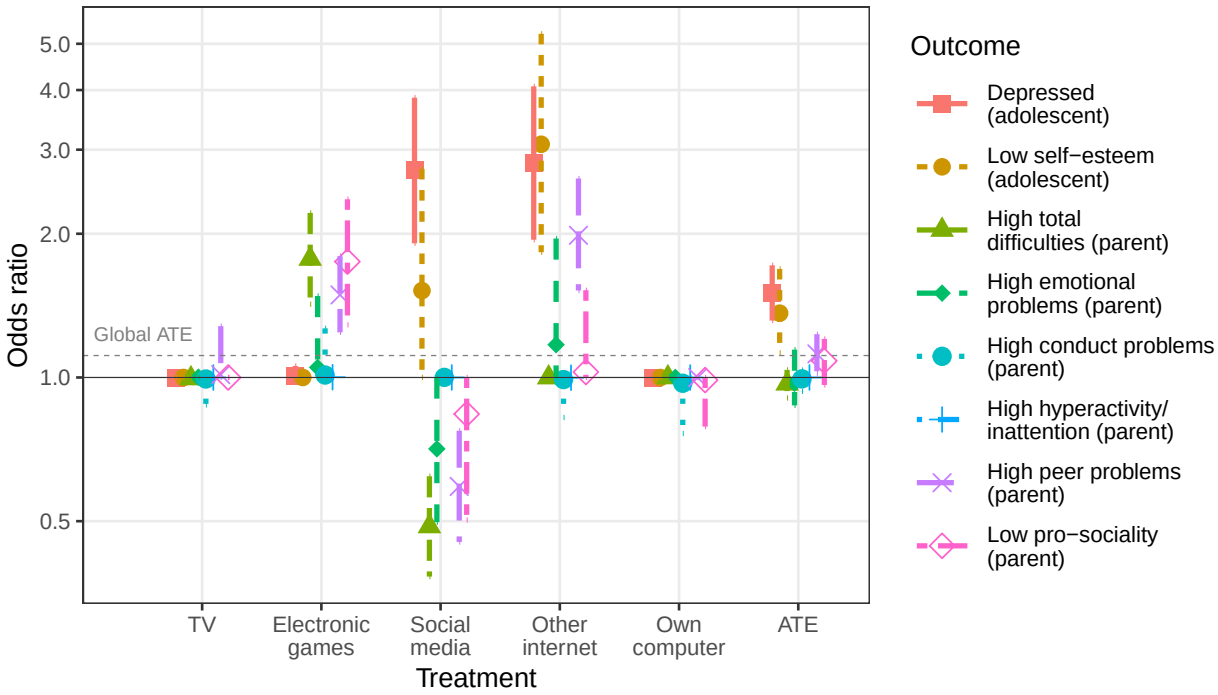
3.4 Reproducing Figure 2 (multiple-outcome BSCA)

Our MCS data analysis included eight outcomes and five treatments of interest, for a total of 40 treatment-outcome combinations. To reduce the burden associated with producing a single BSCA plot for each outcome, the function `multi_bsca` summarizes the BMA results in a single plot. This plot allows one to easily evaluate and compare effects of several treatments on several outcomes. For instance, below all treatments have a similar effect on adolescent-assessed depression and on self-stem. However, these effects are non-comparable to those on parent-assessed total difficulties and emotional problems.[3] We also add the ATE across treatments (`add.ate=TRUE`) and the simple average across outcomes (`add.avg=TRUE`). The average of the ATE estimates is the global ATE.

```
g = multi_bsca(mcs_coef, ms=mcs_ms, conversion=exp,  y.scale='Odds ratio',
           treatments=c('TV', 'Electronic games', 'Social media',
                        'Other internet',  'Own computer'),
```

[3]Using the `BAS` package (if `fast=FALSE`), currently shows normal approximations for the 95% interval in the summary plot. Thus, it looks different from the `mombf` version, which estimates the 95% interval using posterior sampling. Since the latter are more exact, we show it in the main paper. However, running the above analysis using `BAS` shows that the main results remain unchanged (up to rounding) if one samples the entire model space. Adding the ATE is not supported for the `BAS` package.

```
            add.ate=TRUE, add.global.ate=TRUE) +
  scale_y_continuous(trans='log', breaks=c(1/2,1:5), minor_breaks=NULL) +
  theme(legend.key.size=unit(0.9, 'cm'))
g + geom_hline(yintercept=1, lwd=0.2, colour=g$theme$panel.border$colour)
```



4 Robustness checks

In this section, we provide robustness checks for the results presented in the main paper.

4.1 YRBS: alternative outcomes

In their analysis of the YRBS data, besides 'thought about suicide', Orben and Przybylski (2019) used several well-being measures to study the effect of technology use: loneliness, planned suicide, attempted to commit suicide and saw a doctor about suicide. We have reproduced Fig. 1a for these outcomes in Fig. S1. All associations are qualitatively similar, although the magnitudes vary.

4.2 YRBS: linear regression

All outcome variables in the YRBS are binary (e.g. 0 = did not think about suicide, 1 = thought about suicide). Orben and Przybylski (2019) used linear regression, which is unsuitable for binary outcomes. Instead, we used logistic regression to model the probability of the outcome being one. The logistic regression coefficient is easily interpretable; it is the log odds-ratio associated of said probability relative to the reference group, for example the

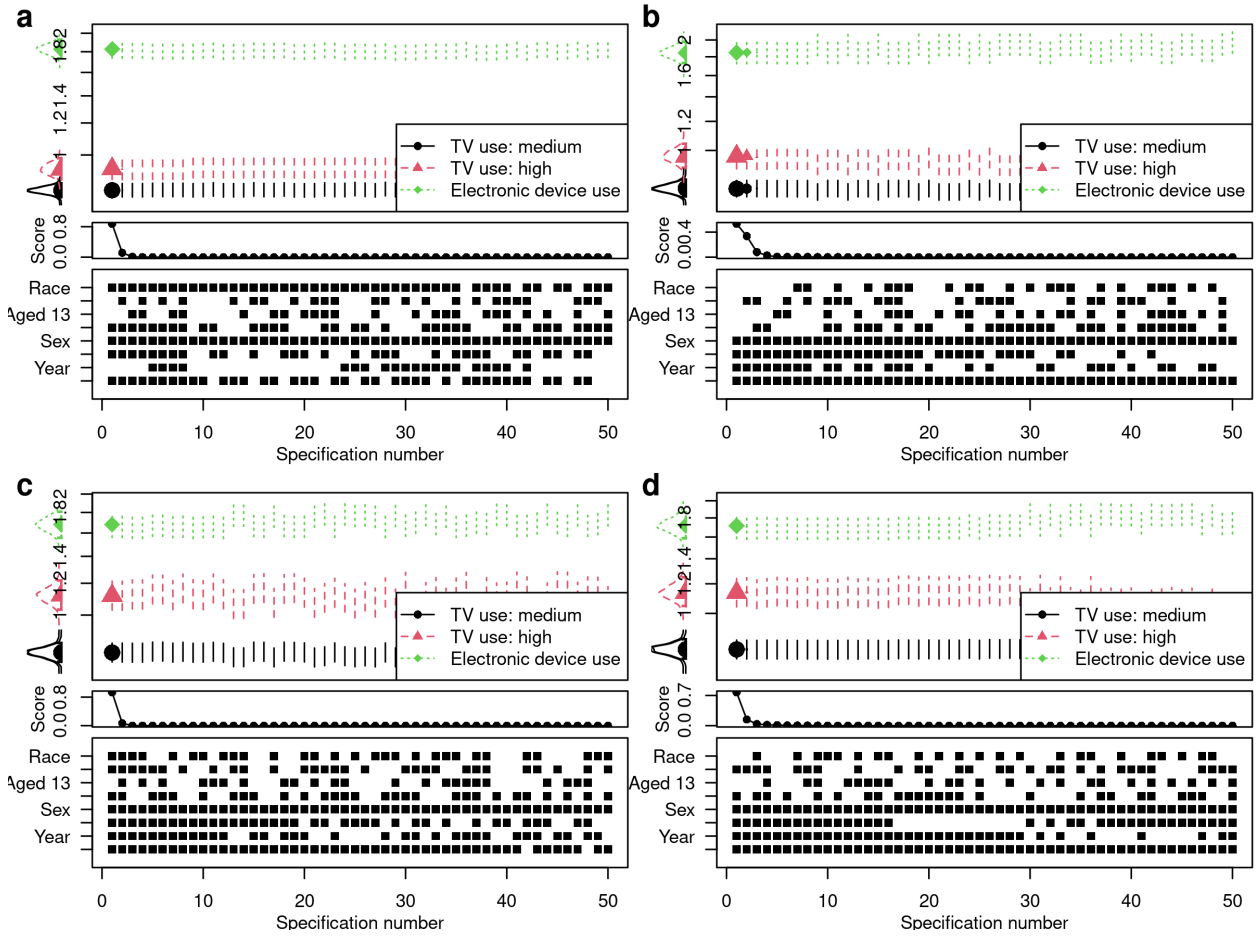


Figure S1: BSCA for other outcomes using YRBS data: (a) loneliness, (b) planned suicide, (c) attempted to commit suicide and (d) saw a doctor about suicide. Otherwise, everything – including the order of control variables in the bottom panel – is as in Fig. 1a of the main text.

log odds-ratio of the probability that a teenager who watches a moderate amount of TV thinking about suicide relative to a teenager who does not watch TV.

For robustness, Fig. S2 shows the single-outcome BSCA for a linear regression model. Our main results remain qualitatively unaltered: electronic device use is associated with an increase in the probability of thinking about suicide, whereas moderate TV use is associated with a decrease.

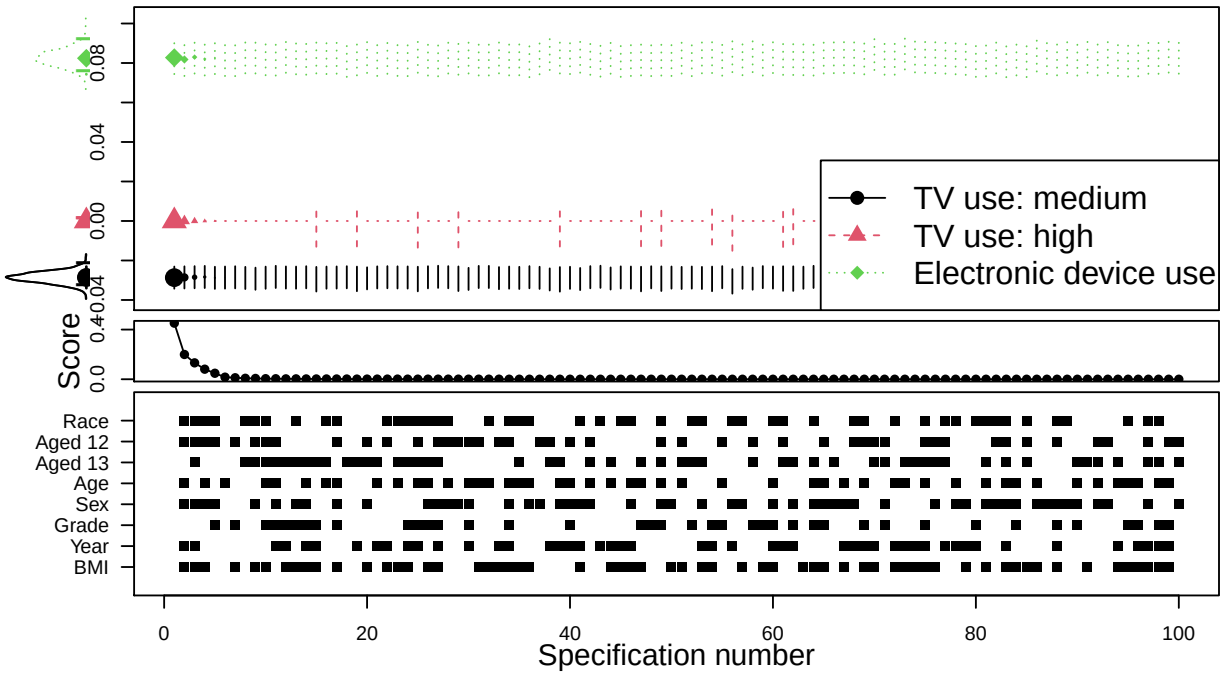


Figure S2: YRBS data. BSCA for the effect technology uses on thinking about suicide. Estimates are obtained using the linear probability model, by iterating over all possible models – including with only one TV coefficient removed. Otherwise, everything is specified as in Fig. 1a of the main text.

4.3 YRBS: lineary of association with TV use

Preliminary exploratory data analyses (see the supplementary file `bsca_preanalysis.html`) revealed that in both the YRBRS and MCS datasets almost all treatments had a monotone, near-linear association. The only exception was TV usage in the YRBS data, which displayed a U-shaped association with adolescent well-being, see Fig. S3.

Therefore, we estimated separate coefficients for medium and high TV use based on coefficient similarity in Fig. S3 (see the main text for the exact coding). For completeness, we also performed a BSCA with an assumed linear TV effect. As expected, the association was in between that for medium and high usage, as shown in Fig. S4. Note that the estimate still suggests that higher TV use is associated with lower odds of thinking about suicide, as in our main analysis.

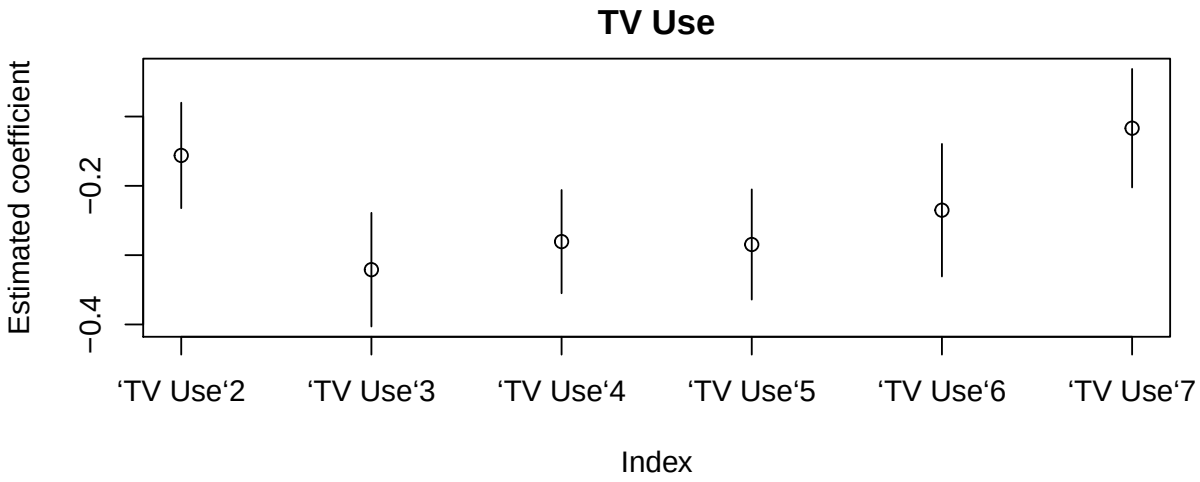


Figure S3: YRBS data. Estimated coefficient for different values of TV use on thinking about suicide. Coefficients were estimated by MLE for a model including all treatment and control variables. Estimated coefficients are qualitatively similar for other outcomes.

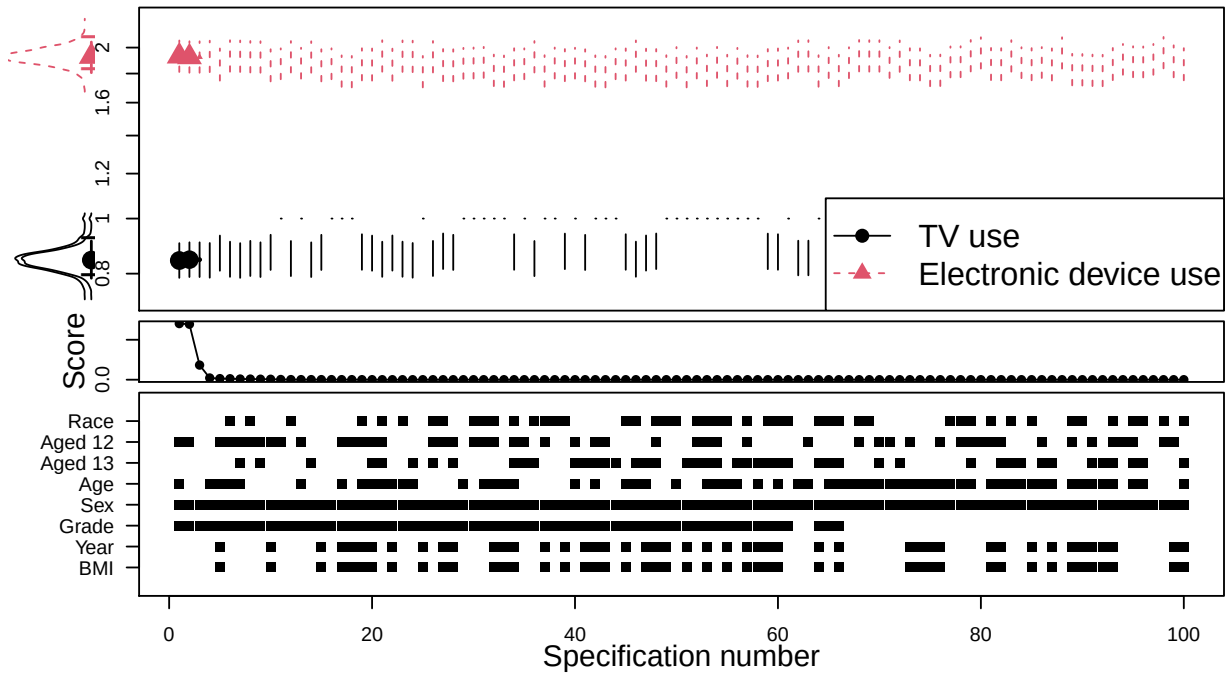


Figure S4: YRBS data. Bayesian SCA for the effect of technology uses on thinking about suicide. TV use is numeric, with 0 = no usage and 1 = 5 hours or more. Otherwise, everything is specified as in Fig. 1a of the main text.

4.4 MCS: full BSCAs

Fig. 2 of the main paper shows the summary of coefficients for different (parent- and self-assessed) outcomes. We showed the multiple-outcome BSCA for brevity. Fig. S5 shows the single-outcome BSCAs for those outcomes discussed in the text (self-assessed depression and low self-esteem, parent-assessed high total difficulties and emotional problems), whereas Fig. S6 shows those for the remaining parent-assessed outcomes.

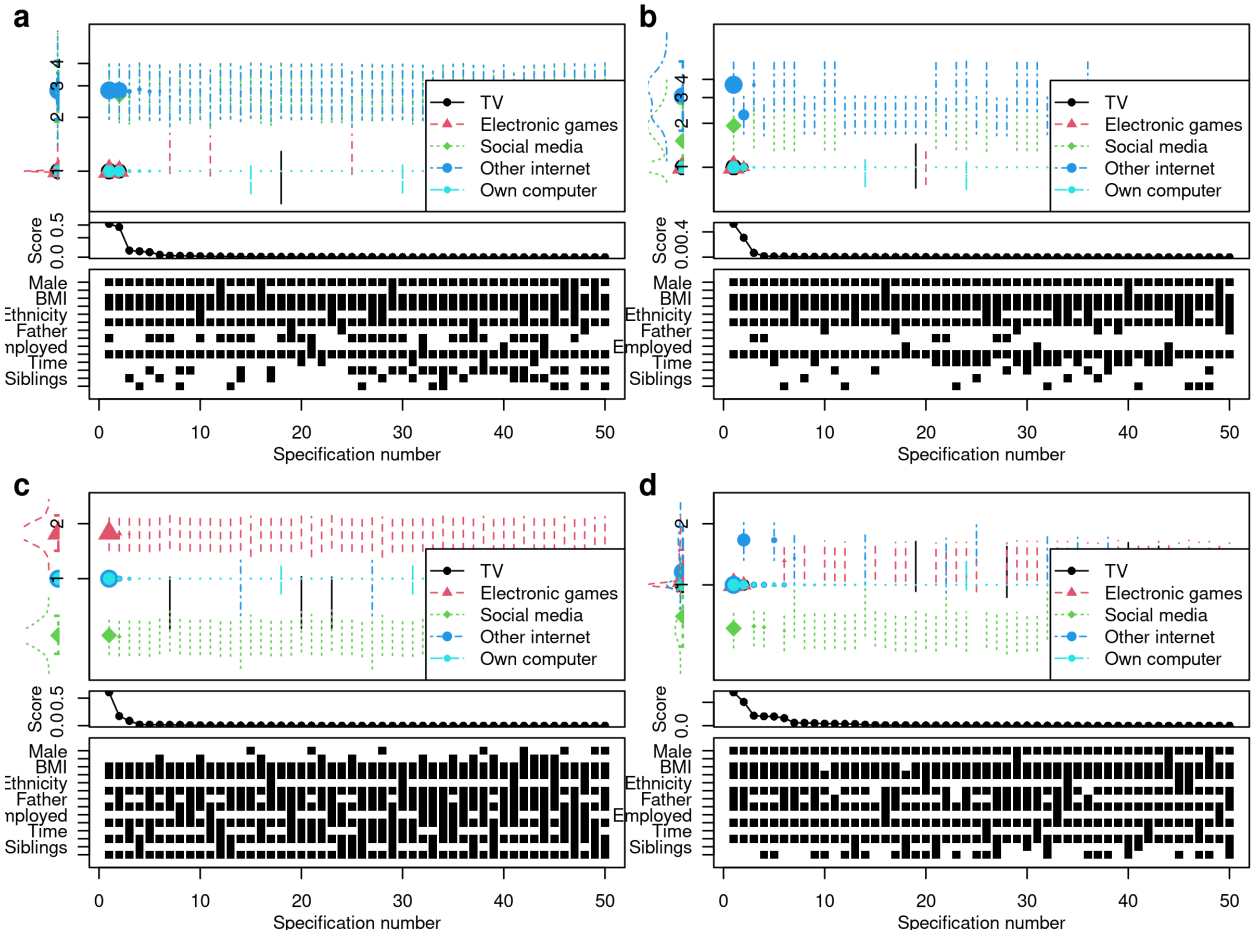


Figure S5: MCS data. BSCAs for outcomes: (a) adolescent-assessed depression, (b) adolescent-assessed low self-esteem, (c) parent-assessed high total difficulties and (d) parent-assessed high emotional problems. Otherwise, everything – including the order of control variables in the bottom panel – is as in Fig. 1b of the main text. For variable codings, see Fig. 2 of the main text.

4.5 MCS: linear regression

Unlike in the YRBS dataset, which has binary outcomes, the outcomes in the MCS dataset are numerical. Parent assessments use the Strengths and Difficulties Questionnaire (SDQ),

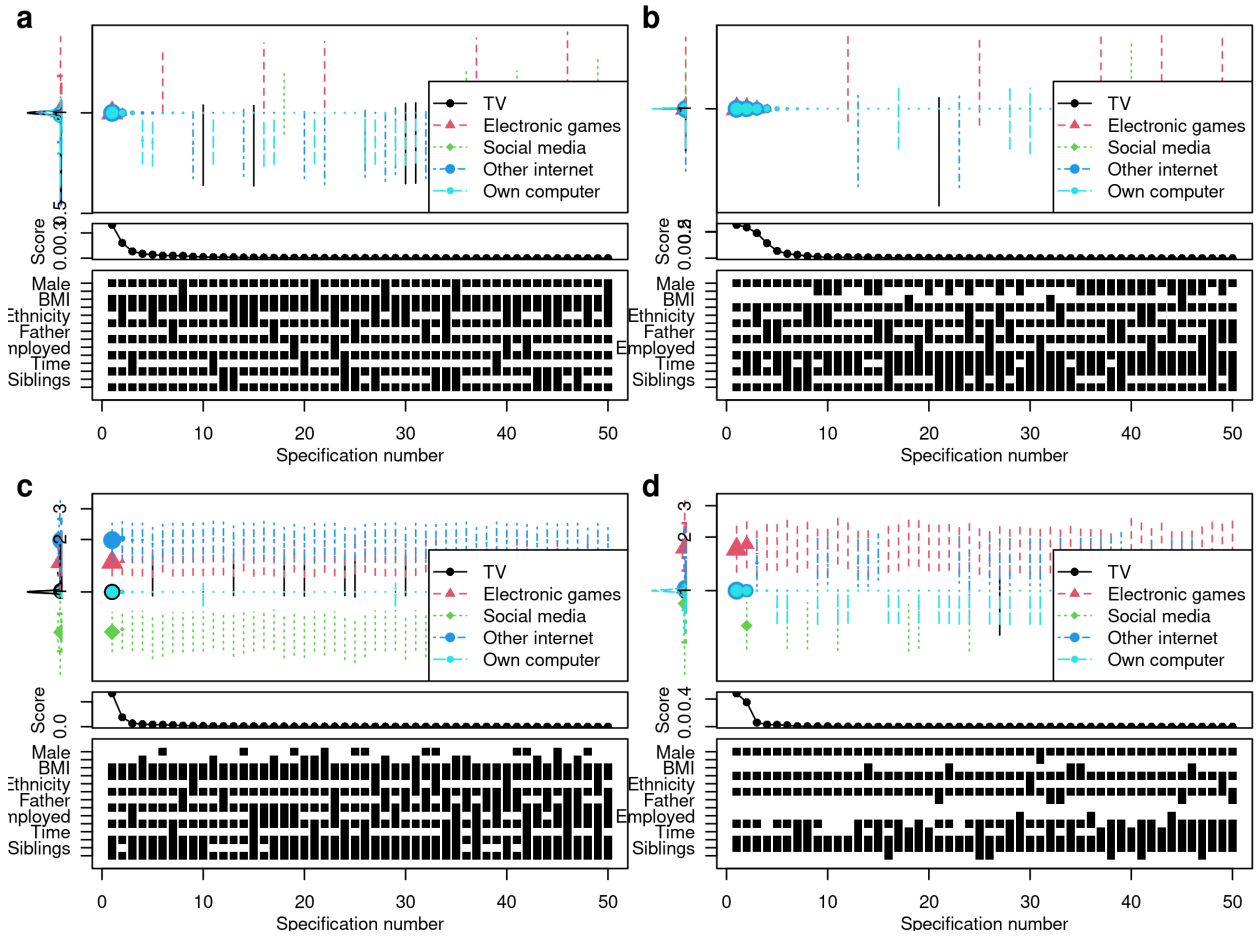


Figure S6: MCS data. BSCAs for parent-assessed outcomes (SDQ): (a) conduct problems (b) hyperactivity/inattention, (c) peer problems and (d) prosociality. Otherwise, everything – including the order of control variables in the bottom panel – is as in Fig. 1b of the main text. For variable codings, see Fig. 2 of the main text.

which provides a score for several well-being categories. Adolescent self-assessments come in three different forms: depressive symptoms according to the short version of the Mood and Feelings Questionnaire (SMFQ), self-esteem according to the Rosenberg scale, and a non-standardized "wellbeing grid" (with the prompt "On a scale of 1 to 7 where '1' means completely happy and '7' means not at all happy, how do you feel about the following parts of your life?" and various outcomes, e.g. 'Your life as a whole'). Individual questions use 3 item (SDQ and SMFQ), 4 item (Rosenberg) or 7 item (wellbeing grid) Likert scales and the aggregate scores (where available) also have different supports.

In our main analysis we converted the outcomes into a binary outcome where 1 codes for an abnormal outcome and 0 for a normal outcome. This conversion was done for two reasons. First, we could use logistic regression for both the YRBS and MCS data, making the results (i.e. odds ratios between low and high usage) comparable. Second, because the scales are mostly standardized, there exist recommended cutoffs derived from population data to define our binary bad (abnormal) / good (normal) outcome. The cutoffs (given in Fig. 2 of the main text) come from Goodman (1997) and Goodman, Meltzer, and Bailey (1998) for the SDQ, Rosenberg (1965) and Nguyen et al. (2019) for the Rosenberg scale, and Thabrew et al. (2018) for the SMFQ. The 'well-being grid' outcomes are not included in Fig. 2, since they do not come from standardized scales and no population-based abnormality cutoffs are available, see next section.

For robustness, we also obtained results using the linear regression model on the original numerical outcomes. The linear regression results for the outcomes discussed in the main paper are shown in Fig. S7. One important difference is that social media is not included in the results for the self-esteem outcome. This is likely due to the high co-linearity with other internet (Pearson's correlation = -0.63). Otherwise, our main results retain their estimated signs.

In addition, Figs. S8 and S9 show the linear results for the remaining parent-assessed and adolescent-assessed outcomes, respectively. Notably, unlike in the logistical model, in Fig. S8 other (non-social media) internet usage and owning a computer are negatively and positively associated with parent-assessed prosociality. As shown in Fig. S9, technology use has very different estimated effects on different types of self-assessed happiness. Social media, for example, is positively associated with being happy with ones friends, but negatively with being happy with ones looks. In contrast, none of the technology uses considered are associated with happiness regarding school, school work or family.

5 Further differences with Orben and Przybylski (2019)

Our analysis is based on the YRBS and MCS datasets also used by Orben and Przybylski (2019) and largely followed their treatment of the data, which was facilitated by their commendable sharing of the R code used for the analysis. However, we deviated from their data treatment choices (in the MCS data) when we felt these were potentially problematic. We also enlarged the set of possible control variables in both datasets. In this section, we explain these differences.

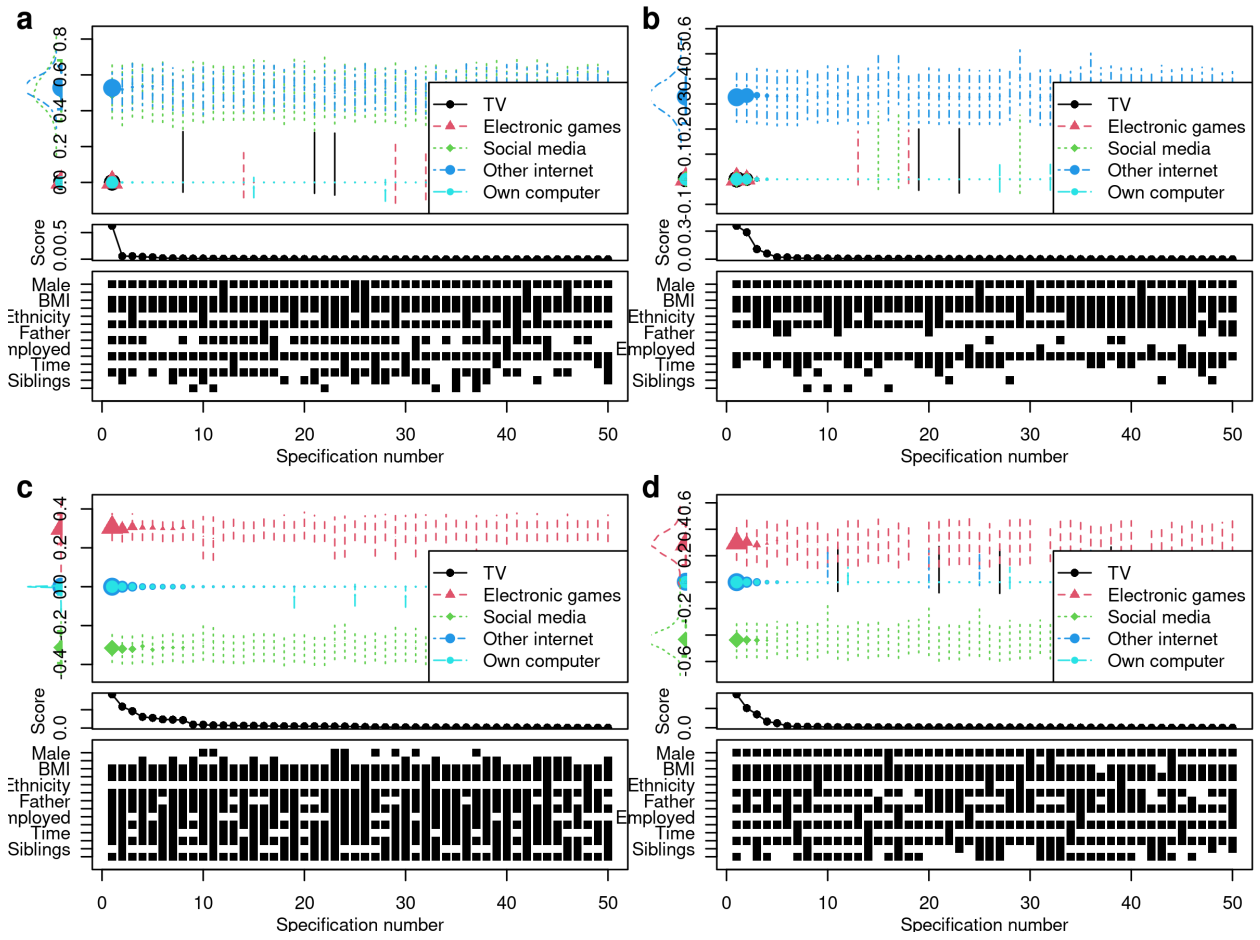


Figure S7: MCS data. Linear regression BSCA for main outcomes: (a) adolescent-assessed depression, (b) adolescent-assessed self-esteem (inverted), (c) parent-assessed total difficulties and (d) parent-assessed emotional problems.

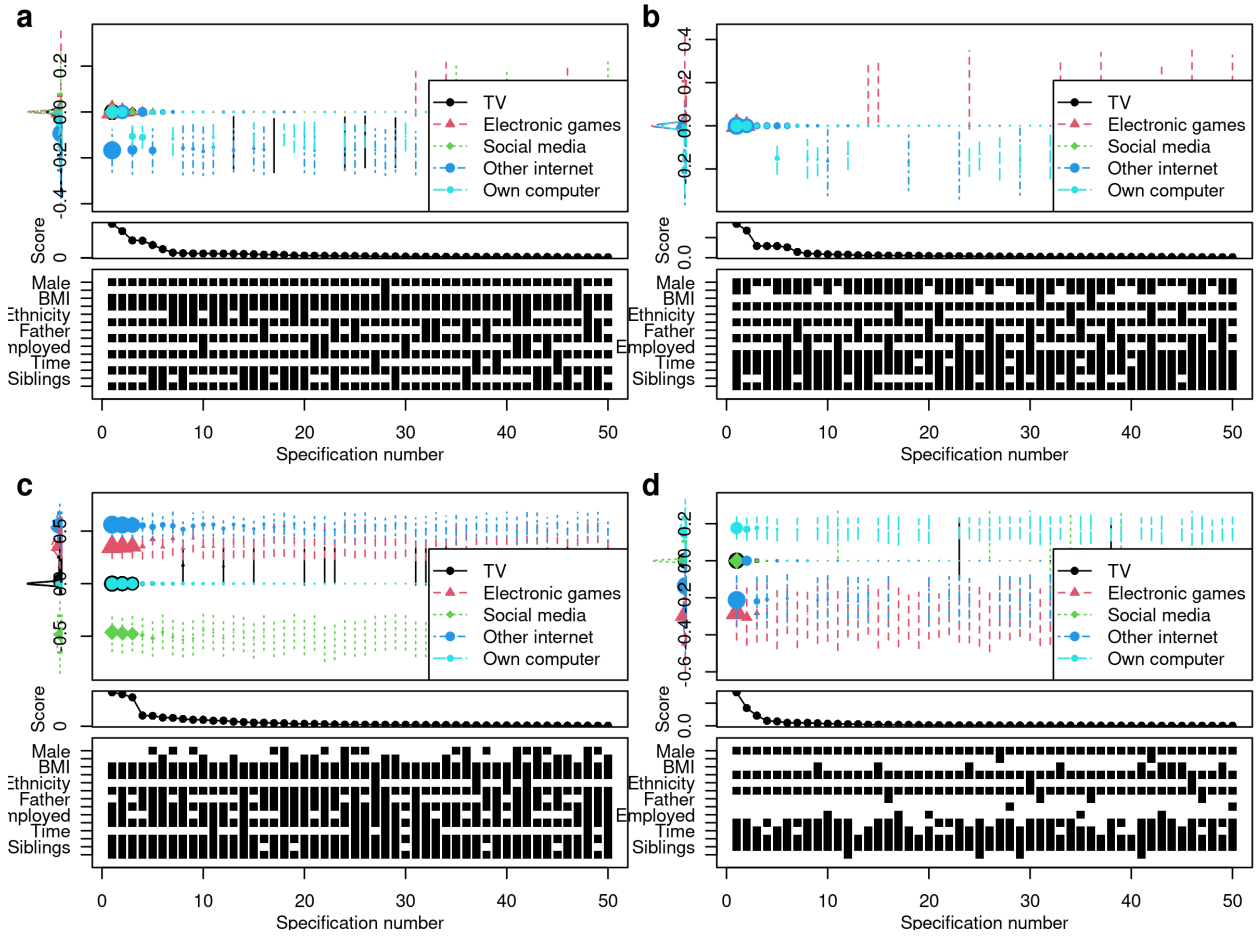


Figure S8: MCS data. Linear regression BSCA for parent-assessed outcomes (SDQ): (a) conduct problems (b) hyperactivity/inattention, (c) peer problems and (d) prosociality.

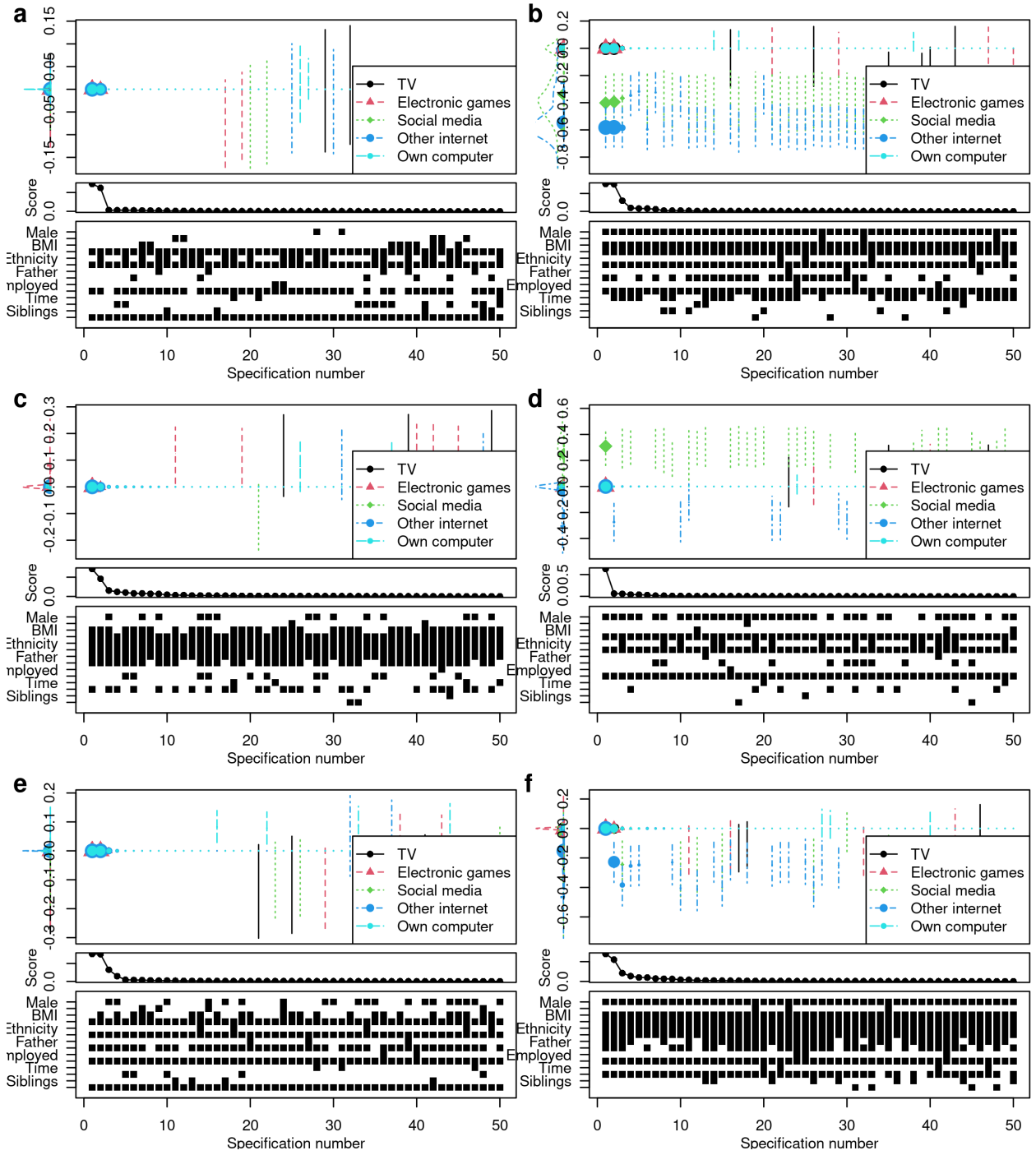


Figure S9: MCS data. Linear regression BSCA for adolescent-assessed outcomes (well being grid): happy with (a) school work (b) looks, (c) family, (d) friends, (e) school and (f) life as a whole.

Table S1: Summary statistics for parent-assessed MCS outcome variables used in Orben and Przybylski (2019)

	Obs.	Mean	S.D.	Min	25%	Median	75%	Max
fpsdro00	10265	0.53	0.68	-1	0	1	1	1
fpsdhs00	10214	0.54	0.65	-1	0	1	1	1
fpsdtt00	10242	0.42	0.71	-1	0	1	1	1
fpsdsp00	10234	0.48	0.68	-1	0	1	1	1
fpsdmw00	10242	0.50	0.66	-1	0	1	1	1
fpsdfs00	10198	0.67	0.60	-1	0	1	1	1
fpsdfb00	10255	0.92	0.32	-1	1	1	1	1
fpsdud00	10221	0.74	0.54	-1	1	1	1	1
fpsddc00	10228	0.32	0.72	-1	0	0	1	1
fpsdnc00	10240	0.52	0.66	-1	0	1	1	1
fpsdoa00	10240	0.80	0.46	-1	1	1	1	1
fpsdpb00	10220	0.71	0.57	-1	1	1	1	1
fpsdcs00	10249	0.94	0.29	-1	1	1	1	1
fpsdgb00	10249	0.46	0.67	-1	0	1	1	1
fpsdfe00	10271	0.68	0.57	-1	0	1	1	1
fconduct	11488	9.58	1.63	1	9	10	11	11
fhyper	11481	8.01	2.40	1	7	8	10	11
fprosoc	11489	8.31	1.85	0	7	9	10	10
fpeer	11491	9.26	1.82	1	8	10	11	11
femotion	11486	8.95	2.14	1	8	10	11	11
febdtot	11471	32.81	5.99	3	30	34	37	41

5.1 Re-scaling of variables

Orben and Przybylski (2019) transformed all outcome variables into a common 1-10 point scale prior to their analysis, so that it is easier to compare outcome values, estimated treatment effects, and combining them across outcomes. Unfortunately, this pre-processing step was not coherently applied to all MCS outcomes (see `1_3_prep_mcs.R` in their replication files), which led to several outcomes not being in the 1-10 scale (some actually have negative values, whereas one outcome takes on values as large as 41, see Table S1). As a consequence, the estimated treatment effects for these outcomes are not really comparable to the outcomes that were in the 1-10 scale, leading to difficulties in interpreting their SCA plot.

5.2 Questionary outcomes: individual questions versus validated scales

The analysis by Orben and Przybylski (2019) used as outcomes individual questions that make up common scales (e.g. all the values ending in 00 in Table S1). This is done for the

adolescent-assessed Mood and Feelings Questionnaire – short version (SMFQ), the adolescent-assessed Rosenberg scale and the parent-assessed Strengths and Difficulties Questionnaire (SDQ). While we think it is generally useful to look at different outcomes, we do not recommend breaking up questions that make up established scales, unless there is a strong reason for doing so. This is because the combined scores have well-established psychometric properties, such as internal consistency, test-retest reliability and validity (see Stone et al. 2010; Sinclair et al. 2010; Thabrew et al. 2018 for the SDQ, Rosenberg scale and SMFQ, respectively). Hence, our analyses focused on the combined scores, rather than on individual questions. This point was also addressed by Orben and Przybylski (2020) themselves.

5.3 Multiple treatment variables

As described in Section 1, in situations where there are multiple treatment variables it is statistically preferable to include them jointly in the model, to ameliorate the confounding between their estimated effects. The YRBS data has two treatments: TV and electronic devise use. The MCS data has five treatments: TV, electronic games, social media, other internet and own computer. In our analyses we always jointly included all treatments.

5.4 Control variables

In our analysis, we included more control variables than Orben and Przybylski (2019). For the YRBS analysis, the only control variable they sometimes included is race. We added age, sex, grade, year of the survey and body-mass index (BMI), several of which were statistically significant (see Fig. 1a and Fig. S1). For the MCS analysis, Orben and Przybylski (2019) included a larger set of control variables (see Figs. 1b and 2), to which we added sex, age and BMI, which again turned out to be important. We also departed from their analysis in how we treated two MCS control covariates: primary caretaker's employment and education status.

Regarding employment, Orben and Przybylski (2019) used the NS-SEC 5 category for the current job (1="Manager" through 5="Routine"), meaning that all kids with unemployed primary caretakers (for which the variable is coded `NA`) are excluded from the analysis. This is actually a non-negligible proportion of the dataset, see Table S2. We believe that such exclusion may introduce biases, by restricting the scope of the inference to kids with employed parents caretakers. Instead, in our analysis we included a binary variable to control for employment status (1=employed or self-employed; 0 otherwise), which allows including kids with unemployed parents into the analysis.

Regarding education status, the covariate is also coded on a 1-5 scale, but there are two special values (95= "Overseas qualification" and 96="None of these"), see Table S3. Orben and Przybylski (2019) included the 95-96 values in their linear regression analysis, which is inappropriate: one cannot interpret the estimated coefficients as capturing the association between the outcome and the caretaker's education status. Possible alternative analyses are to either exclude individuals with these codes (excluding a non-negligible proportion of individuals), or use a non-linear coding for the covariate's effect. For simplicity, and given that we already included numerous other control covariates, in our analysis we excluded the education variable.

Table S2: Primary caretaker employment status (1-5). Number of individuals with each value

1	2	3	4	5	NA
3276	1832	696	273	2012	3795

Table S3: Primary caretaker education status (1-5). Number of individuals with each value

1	2	3	4	5	95	96	NA
660	2598	1496	3784	1170	331	1104	741

References

Benjamini, Y., and Y. Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society B* 57 (1): 289–300.

Castillo, I., J. Schmidt-Hieber, and A. W. van der Vaart. 2015. "Bayesian Linear Regression with Sparse Priors." *The Annals of Statistics* 43 (5): 1986–2018.

Chen, J., and Z. Chen. 2008. "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces." *Biometrika* 95 (3): 759–71.

Clyde, Merlise, Mine Cetinkaya-Rundel, Colin Rundel, David Banks, Christine Chai, and Lizzy Huang. 2020. *An Introduction to Bayesian Thinking*. https://statswithr.github.io/book/.

Dawid, A. P. 1999. "The Trouble with Bayes Factors." University College London.

Efron, Bradley. 2007. "Size, Power and False Discovery Rates." *The Annals of Statistics* 35 (4): 1351–77.

Friel, Nial, and Jason Wyse. 2012. "Estimating the Evidence–a Review." *Statistica Neerlandica* 66 (3): 288–308.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. 3rd ed. Boca Raton: Chapman and Hall/CRC. http://www.stat.columbia.edu/~gelman/book/BDA3.pdf.

Goodman, R, H Meltzer, and V Bailey. 1998. "The Strengths and Difficulties Questionnaire: A Pilot Study on the Validity of the Self-Report Version." *Adolescent Psychiatry* 7 (3): 6.

Goodman, Robert. 1997. "The Strengths and Difficulties Questionnaire: A Research Note." *Journal of Child Psychology and Psychiatry* 38 (5): 581–86. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14: 382–401.

Johnson, V. E., and D. Rossell. 2010. "On the Use of Non-Local Prior Densities for Default Bayesian Hypothesis Tests." *Journal of the Royal Statistical Society B* 72: 143–70.

———. 2012. "Bayesian Model Selection in High-Dimensional Settings." *Journal of the*

American Statistical Association 24 (498): 649–60.

Kass, R. E., L. Tierney, and J. B. Kadane. 1990. "The Validity of Posterior Expansions Based on Laplace's Method." *Bayesian and Likelihood Methods in Statistics and Econometrics* 7: 473–88.

Kass, Robert E, and Adrian E Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–95.

LeCam, Lucien. 1953. "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates." *Univ. California Pub. Statist.* 1: 277–330.

Madigan, D., and A. E. Raftery. 1994. "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window." *Journal of the American Statistical Association* 89 (428): 1535–46.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. 2nd ed. Boca Raton: Chapman and Hall/CRC.

Nguyen, Dat Tan, E. Pamela Wright, Christine Dedding, Tam Thi Pham, and Joske Bunders. 2019. "Low Self-Esteem and Its Association with Anxiety, Depression, and Suicidal Ideation in Vietnamese Secondary School Students: A Cross-Sectional Study." *Frontiers in Psychiatry* 10. https://doi.org/10.3389/fpsyt.2019.00698.

Orben, Amy, and Andrew K. Przybylski. 2019. "The Association Between Adolescent Well-Being and Digital Technology Use." *Nature Human Behaviour* 3 (2): 173–82. https://doi.org/10.1038/s41562-018-0506-1.

———. 2020. "Reply to: Underestimating Digital Media Harm." *Nature Human Behaviour* 4 (4): 349–51. https://doi.org/10.1038/s41562-020-0840-y.

Rosenberg, Morris. 1965. *Society and the Adolescent Self-Image*. Princeton University Press.

Rossell, D. 2018. "A Framework for Posterior Consistency in Model Selection." *arXiv* 1806.04071: 1–58.

Rossell, D., and D. Telesca. 2017. "Non-Local Priors for High-Dimensional Estimation." *Journal of the American Statistical Association* 112: 254–65.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics* 6: 461–64.

Scott, J. G., and J. O Berger. 2010. "Bayes and Empirical Bayes Multiplicity Adjustment in the Variable Selection Problem." *The Annals of Statistics* 38 (5): 2587–2619.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour*, July. https://doi.org/10.1038/s41562-020-0912-z.

Sinclair, Samuel J., Mark A. Blais, David A. Gansler, Elisabeth Sandberg, Kimberly Bistis, and Alice LoCicero. 2010. "Psychometric Properties of the Rosenberg Self-Esteem Scale: Overall and Across Demographic Groups Living Within the United States." *Evaluation & the Health Professions* 33 (1): 56–80. https://doi.org/10.1177/0163278709356187.

Stone, Lisanne L., Roy Otten, Rutger C. M. E. Engels, Ad A. Vermulst, and Jan M. A. M. Janssens. 2010. "Psychometric Properties of the Parent and Teacher Versions of the Strengths and Difficulties Questionnaire for 4- to 12-Year-Olds: A Review." *Clinical Child and Family Psychology Review* 13 (3): 254–74. https://doi.org/10.1007/s10567-010-0071-2.

Thabrew, Hiran, Karolina Stasiak, Lynda-Maree Bavin, Chris Frampton, and Sally Merry. 2018. "Validation of the Mood and Feelings Questionnaire (MFQ) and Short Mood and Feelings Questionnaire (SMFQ) in New Zealand Help-Seeking Adolescents." *International*

Journal of Methods in Psychiatric Research 27 (3): e1610. https://doi.org/10.1002/mpr.1610.

Vaart, A. W. van der. 1998. *Asymptotic Statistics*. New York: Cambridge University Press.