# Specification Curve Analysis

JDI OPEN

**Toby Davies**


# Source material

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0506-1

## The association between adolescent well-being and digital technology use

Amy Orben[1*] and Andrew K. Przybylski[1,2]

**The widespread use of digital technologies by young people has spurred speculation that their regular use negatively impacts psychological well-being. Current empirical evidence supporting this idea is largely based on secondary analyses of large-scale social datasets. Though these datasets provide a valuable resource for highly powered investigations, their many variables and observations are often explored with an analytical flexibility that marks small effects as statistically significant, thereby leading to potential false positives and conflicting results. Here we address these methodological challenges by applying specification curve analysis (SCA) across three large-scale social datasets (total $n = 355{,}358$) to rigorously examine correlational evidence for the effects of digital technology on adolescents. The association we find between digital technology use and adolescent well-being is negative but small, explaining at most 0.4% of the variation in well-being. Taking the broader context of the data into account suggests that these effects are too small to warrant policy change.**

RESOURCE

https://doi.org/10.1038/s41562-020-0912-z

nature human behaviour

Check for updates

## Specification curve analysis

Uri Simonsohn[1]✉, Joseph P. Simmons[2] and Leif D. Nelson[3]

**Empirical results hinge on analytical decisions that are defensible, arbitrary and motivated. These decisions probably introduce bias (towards the narrative put forward by the authors), and they certainly involve variability not reflected by standard errors. To address this source of noise and bias, we introduce specification curve analysis, which consists of three steps: (1) identifying the set of theoretically justified, statistically valid and non-redundant specifications; (2) displaying the results graphically, allowing readers to identify consequential specifications decisions; and (3) conducting joint inference across all specifications. We illustrate the use of this technique by applying it to three findings from two different papers, one investigating discrimination based on distinctively Black names, the other investigating the effect of assigning female versus male names to hurricanes. Specification curve analysis reveals that one finding is robust, one is weak and one is not robust at all.**

# The problem

Numerous cases where different studies examining the same underlying research question have conflicting findings. Multiple reasons:

- Natural variation from setting to setting
- Methodological discrepancies
- More nefarious things…

Some of this is due to flaws that we can address by being methodologically *careful* (i.e. eliminating bad practices). But some is due to perfectly legitimate decisions taken by researchers – known as 'researcher degrees of freedom':

- When you have a dataset and a question, you have to make choices in how you analyse
- You and I might make different choices, just because we (honestly) see things in a different way – **legitimate** but **subjective**

# What kind of thing?

Suppose you want to examine whether **people's fear of crime (DV) depends on the physical state of their local environment (IV)**. Suppose you try to use the Crime Survey of England & Wales (CSEW):

- CSEW has hundreds of variables, many measuring similar things
- How do you operationalise DV? Fear of assault, burglary, rape, etc all present – so is 'scared walking down street at night'.
- What about IV? Questions about rubbish in the street, graffiti, run-down homes, etc.
- Which of these should you use? Or do you combine (some of) them? How?
- Many, many defensible ways…

# Specification

This is what is meant by **specification** in this sense – think of it as 'given the data, how do you set up the model to test the thing you want to test'.

- Different decisions lead to different conclusions, even if nobody is doing anything wrong
- Even worse, if someone *is* doing something naughty, they could choose the specification that gives them the desired result

# Existing approaches

Not a new problem – in many papers, multiple models presented, often showing different levels of granularity…

**Table 4** Poisson HLM with random intercepts (*z*-scores shown in parentheses)

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| Level 1 variables | | | |
| Households (hhlds) | 1.03 (66.63)** | 1.03(65.22)** | 1.03 (63.98)** |
| Homes per 100 m | 0.99 (11.41)** | 0.99 (11.10)** | 0.99 (10.47)** |
| Road type | | | |
| Major | 1.22 (5.25)** | 1.11 (2.38)** | 1.10 (2.08)* |
| Minor | 1.25 (5.42)** | 1.24 (4.10)** | 1.21 (3.58)** |
| Private | 0.57 (6.09)** | 0.65 (4.13)** | 0.68 (3.77)** |
| 1st-Order links | 1.03 (2.84)** | – | – |
| Major | – | 1.08 (4.60)** | 1.04 (2.27)* |
| Minor | – | 1.03 (1.74) | 1.00 (0.15) |
| Local | – | 1.02 (2.13)* | 0.99 (0.93) |
| Private | – | 0.91 (2.30)* | 0.87 (3.14)** |
| Cul-de-sac | | | |
| Linear | – | – | 0.91 (2.33)* |
| Sinuous | – | – | 0.73 (7.58)** |

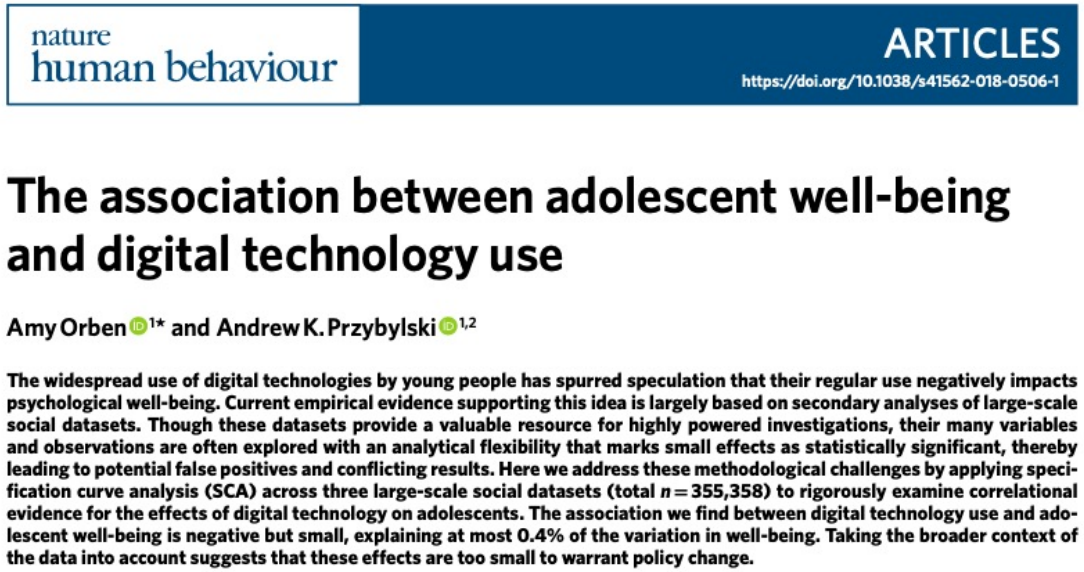
# Specification curve analysis

This approach basically extends and formalises the exploration of multiple models. The basic idea is: **run all legitimate specifications, examine the spectrum of findings, and draw overall conclusions by averaging across them**.

It also involves nice visualisations…

# Orben & Przybylski

A nice demonstration of this idea. They are examining **whether the use of digital technology (phones, laptops, etc) affects well-being in adolescents** – but that is not really the point.



nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0506-1

## The association between adolescent well-being and digital technology use

Amy Orben[1*] and Andrew K. Przybylski[1,2]

**The widespread use of digital technologies by young people has spurred speculation that their regular use negatively impacts psychological well-being. Current empirical evidence supporting this idea is largely based on secondary analyses of large-scale social datasets. Though these datasets provide a valuable resource for highly powered investigations, their many variables and observations are often explored with an analytical flexibility that marks small effects as statistically significant, thereby leading to potential false positives and conflicting results. Here we address these methodological challenges by applying specification curve analysis (SCA) across three large-scale social datasets (total $n = 355{,}358$) to rigorously examine correlational evidence for the effects of digital technology on adolescents. The association we find between digital technology use and adolescent well-being is negative but small, explaining at most 0.4% of the variation in well-being. Taking the broader context of the data into account suggests that these effects are too small to warrant policy change.**

There are large-scale social datasets that measure this kind of thing – they examine three of them, with total *N* of 355,358.

# Models

They try to model the relationship of interest using the simplest possible approach – simple linear regression:

- Try to predict wellbeing (DV) as a function of technology use (IV), as well as potential co-variates (control variables)
- This is deliberately simplistic – again, the complexity of the model is not the point…

# Decisions

**Table 1 | Possible specifications (analytical decisions) used to test a simple linear regression between technology use and adolescent well-being in the datasets YRBS, MTF and MCS**

| Decision | YRBS | MTF | MCS |
|---|---|---|---|
| Operationalizing adolescent well-being | Mean of any possible combination of five items concerning mental health and suicidal ideation | Mean of any possible combination of 13 items concerning depression, happiness and self-esteem | Mean of any possible combination of 24 questions concerning well-being, self-esteem and feelings (cohort members), or mean of any possible combination of 25 questions from the Strengths and Difficulties Questionnaire (caregivers) |
| Operationalizing technology use | Two questions concerning electronic device use and TV use, or the mean of these questions | Eleven technology use measures concerning the Internet, electronic games, mobile phone use, social media use and computer use, or the mean of these questions | Five questions concerning TV use, electronic games, social media use, owning a computer and using the Internet at home, or the mean of these questions |
| Which co-variates to include | Either include co-variates or not | Either include co-variates or not | Either include co-variates or not |
| Other specifications | Either take mean of dichotomous well-being measures, or code all cohort members who answered 'yes' to one or more as 1 and all others as 0 | | Use well-being measures declared by cohort members or those declared by their caregivers |

# Specifications

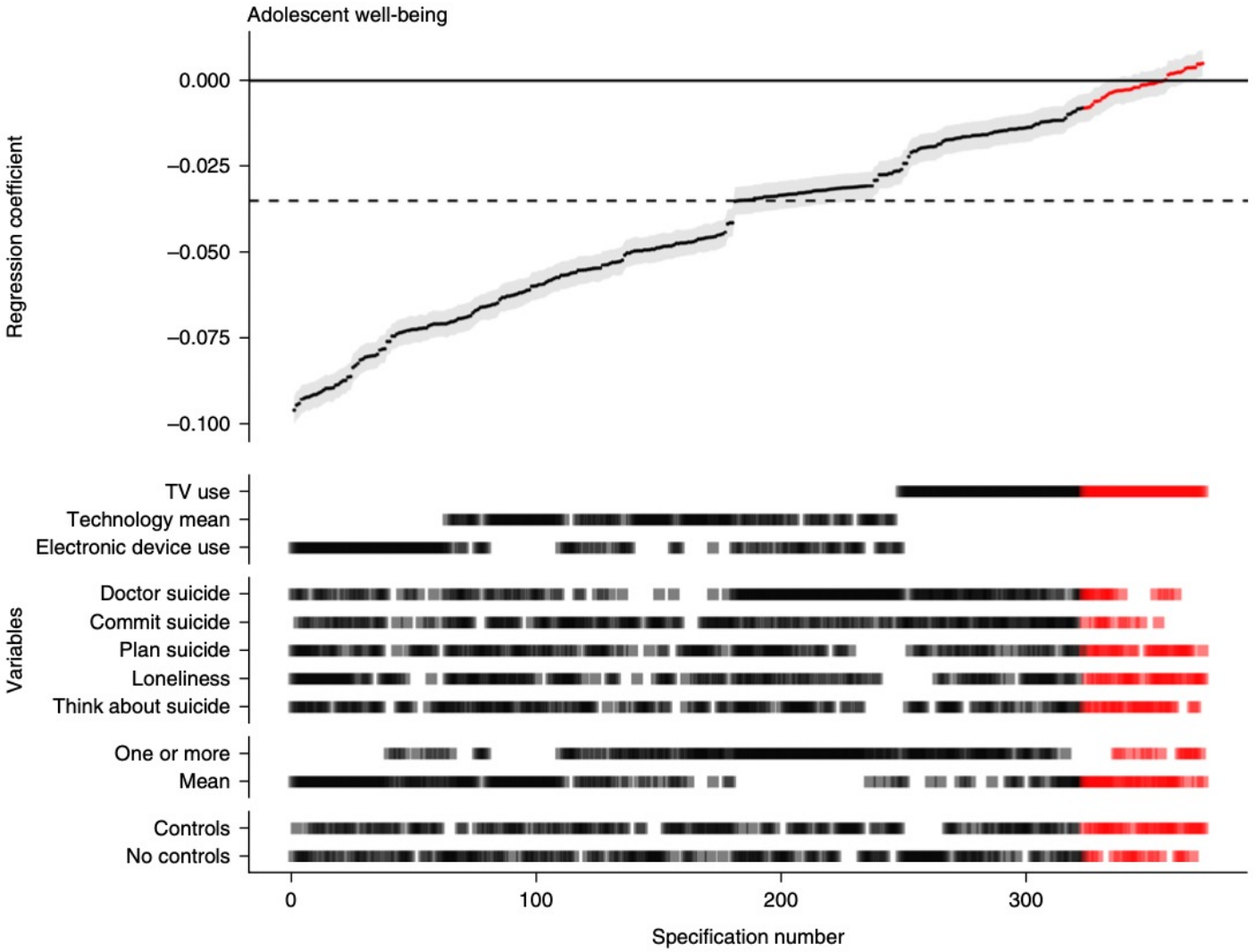
For the 3 main datasets, these decisions lead to:

- 372 plausible specifications for YRBS
- 40,966 plausible specifications for MTF
- 603,979,752 plausible specifications for MCS

So, to generate a specification curve for each dataset:

- Take each possible specification* and run a linear regression
- Note the beta-coefficient (i.e. effect size) in each case
- Plot them all…

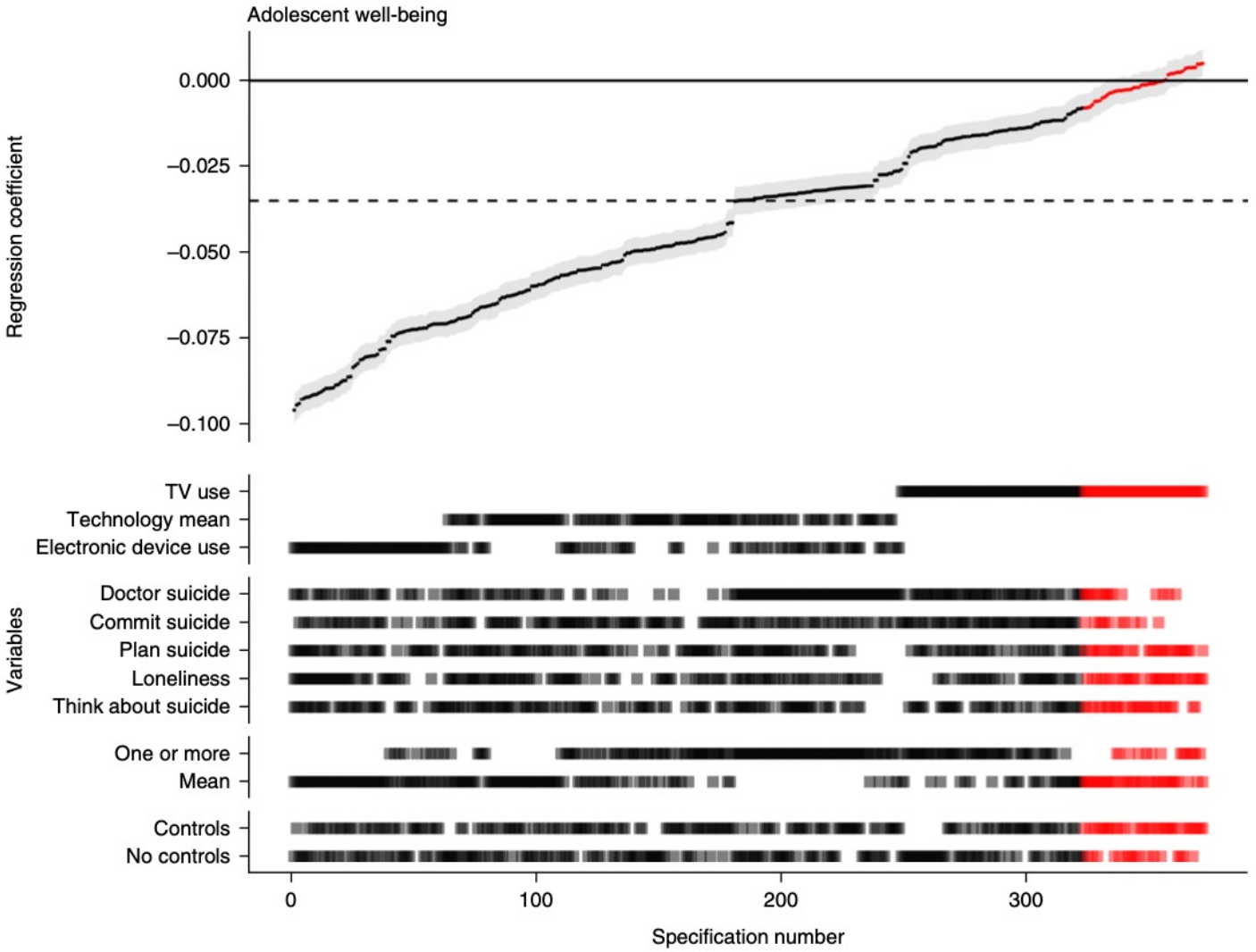
# Specification curve




How to read this:

- Each point on x-axis represents a different specification
- Ordered according to increasing beta-coefficient
- 'Dashboard' below indicates which variables were included in that specification
- Black = significant; red = non-significant
- Dotted line represents median beta-coefficient across all specifications

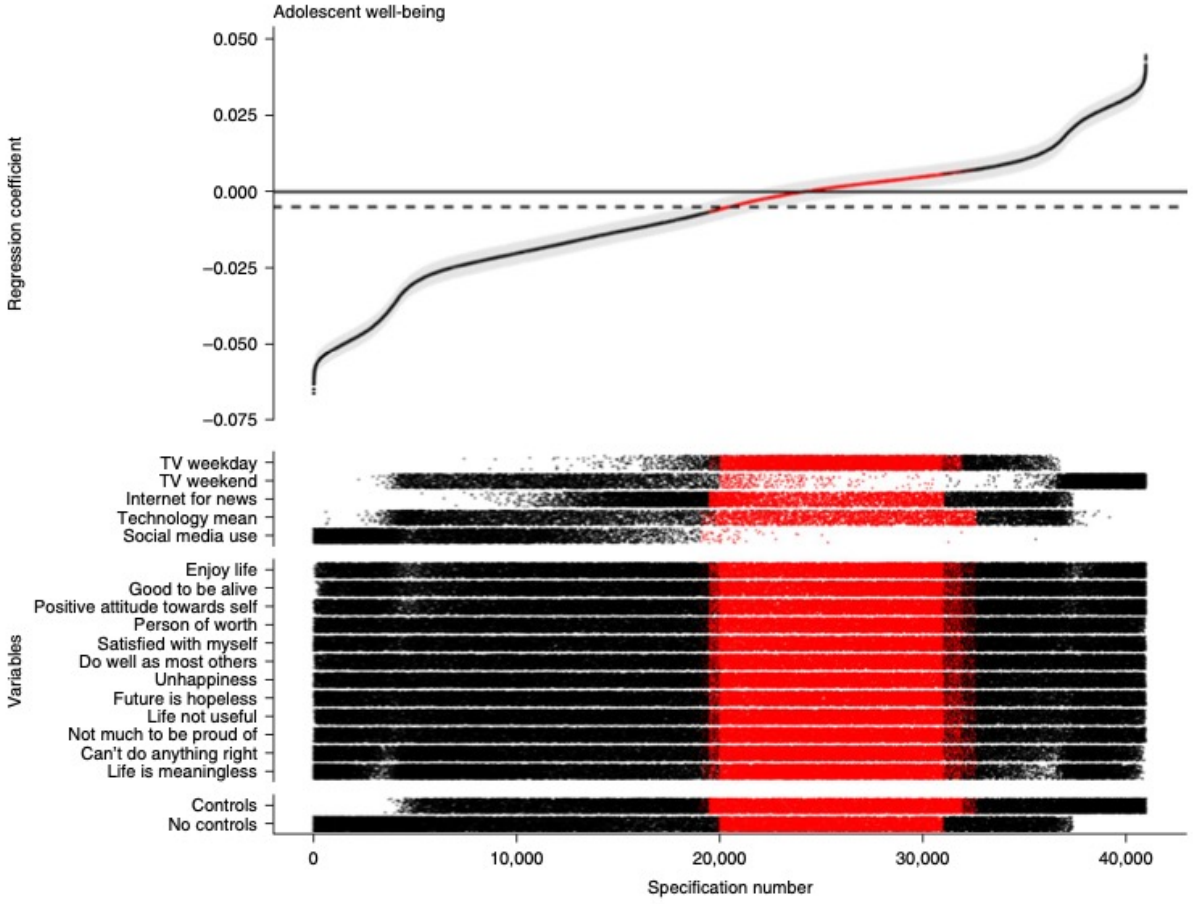
# Specification curve

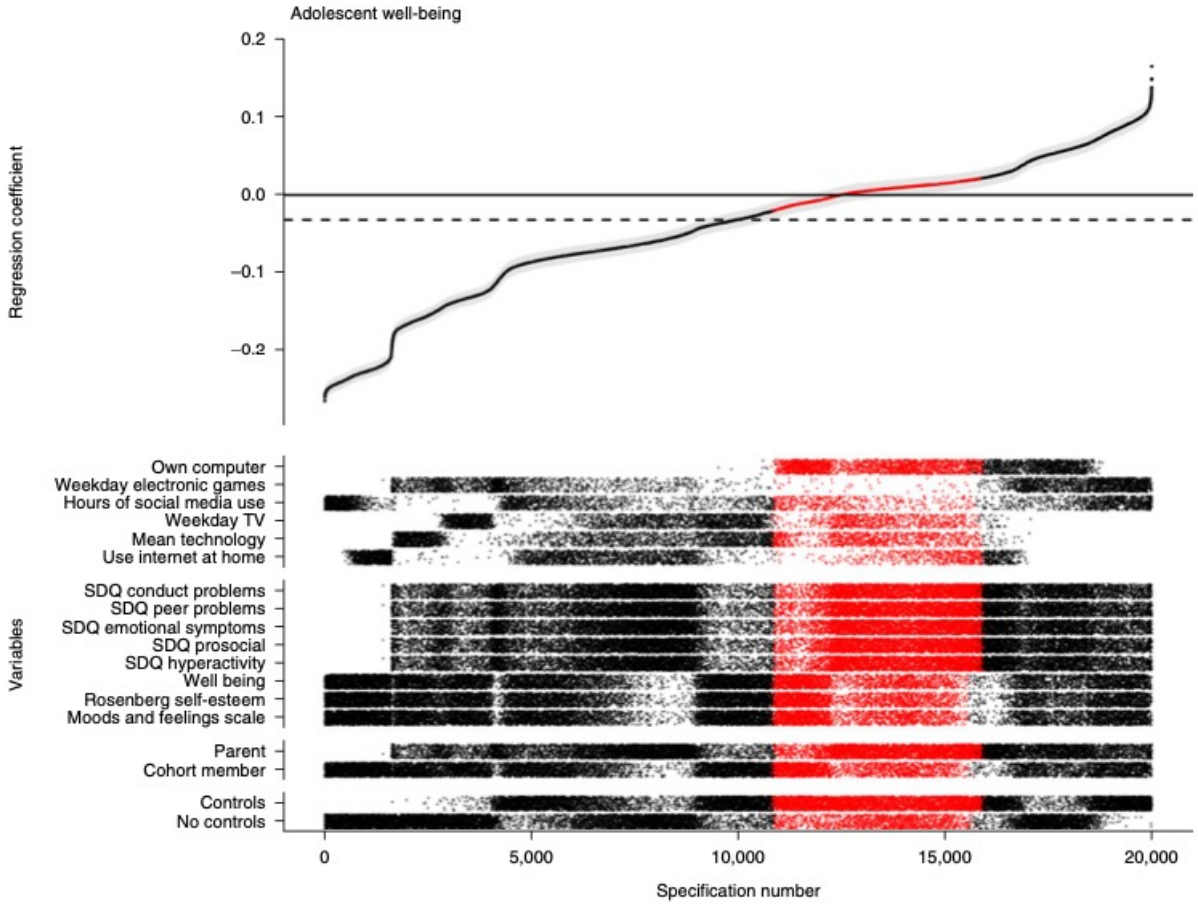



So:

- If my (legitimate, defensible) decisions had led me to a model on the right, I would conclude no effect
- If my decisions had led me to a model on the left, would conclude significant negative relationship
- Most models lead to a conclusion somewhere between those…

# Other datasets



# Thinking numerically

The plots are nice, but can also formally examine the distributions of regression outputs. They examine the medians, and can see the effects of different decisions…

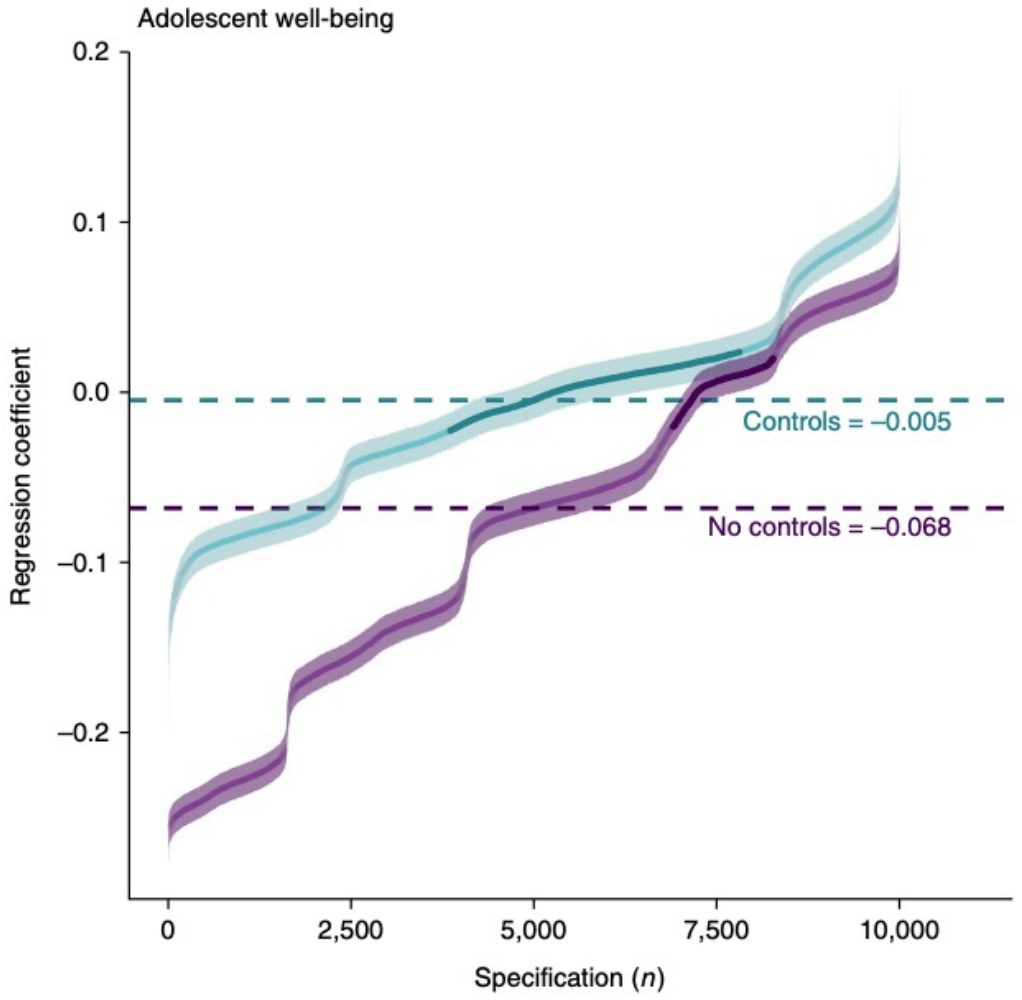
**Table 2 | Results of SCA for the YRBS, MTF and MCS, both overall and for different technology use variables, parent/adolescent self-report or with/without control variables**

| Dataset | Median $\beta$ of SCA | Median partial $\eta^2$ of SCA | Median $n$ | s.e.m. |
|---|---|---|---|---|
| **YRBS** | | | | |
| Complete SCA | −0.035 | 0.001 | 62,297 | 0.004 |
| Electronic device use only | −0.071 | 0.005 | 62,368 | 0.004 |
| TV use only | −0.012 | <0.001 | 62,352 | 0.004 |
| With control variables only | −0.034 | 0.001 | 61,525 | 0.004 |
| Without control variables only | −0.035 | 0.001 | 62,638 | 0.004 |
| **MTF** | | | | |
| Complete SCA | −0.005 | <0.001 | 78,267 | 0.003 |
| Social media use only | −0.031 | 0.001 | 102,963 | 0.003 |
| TV viewing on weekend only | 0.008 | 0.001 | 115,738 | 0.003 |
| Using Internet for news only | −0.002 | <0.001 | 115,580 | 0.003 |
| TV viewing on weekday only | 0.002 | <0.001 | 115,783 | 0.003 |
| With control variables only | 0.001 | <0.001 | 72,525 | 0.003 |
| Without control variables only | −0.013 | <0.001 | 117,560 | 0.003 |
| **MCS** | | | | |
| Complete SCA | −0.032 | 0.004 | 7,968 | 0.010 |
| Own a computer only | −0.003 | 0.011 | 7,973 | 0.010 |
| Weekday electronic games only | 0.013 | <0.001 | 7,977 | 0.010 |
| Hours of social media use only | −0.056 | 0.009 | 7,972 | 0.010 |
| TV viewing on weekday only | −0.043 | 0.003 | 7,971 | 0.010 |
| Use of home Internet only | −0.070 | 0.006 | 7,975 | 0.010 |
| Parent-report well-being only | <0.001 | 0.003 | 7,893 | 0.010 |
| Adolescent-report well-being only | −0.046 | 0.008 | 8,857 | 0.010 |
| With control variables only | −0.005 | 0.001 | 6,566 | 0.011 |
| Without control variables only | −0.068 | 0.005 | 11,018 | 0.010 |

# Thinking numerically

The plots are nice, but can also formally examine the distributions of regression outputs. They examine the medians, and can see the effects of different decisions…

…or can disaggregate the curves themselves.



# Substantive conclusions

For this paper, results taken as a whole suggest a significant negative effect of digital technology on well-being:

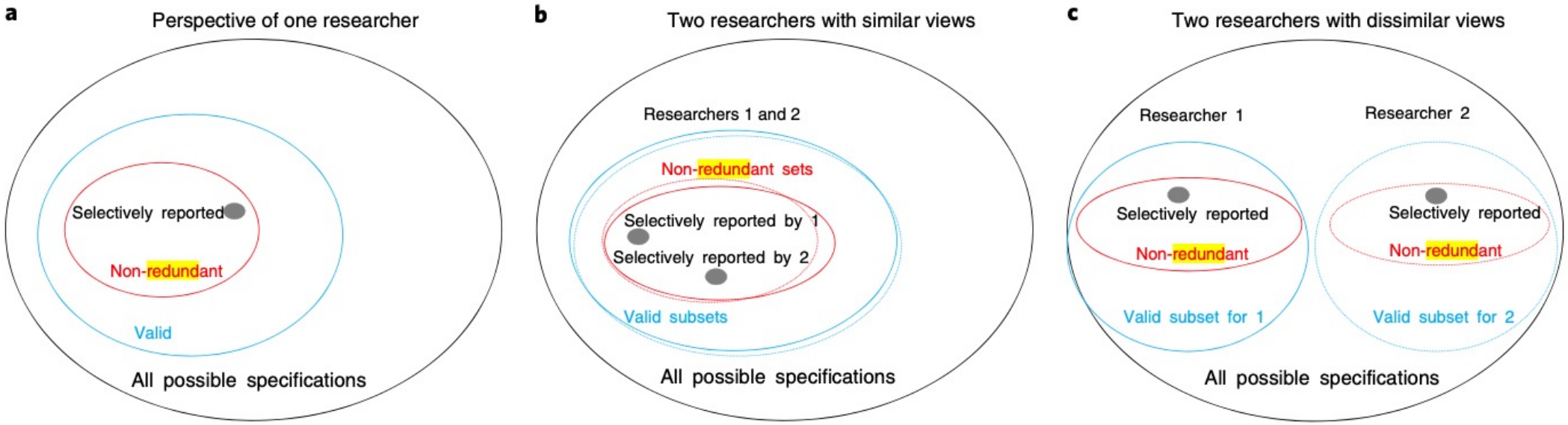
- But the effects are tiny and the variance explained is even less
- They show some entertaining comparisons – effect of digital technology is approximately equal to the effect of regularly eating potatoes, and wearing glasses has 1.5x the effect
- So really not that big an influence…

But again, that wasn't really the point.

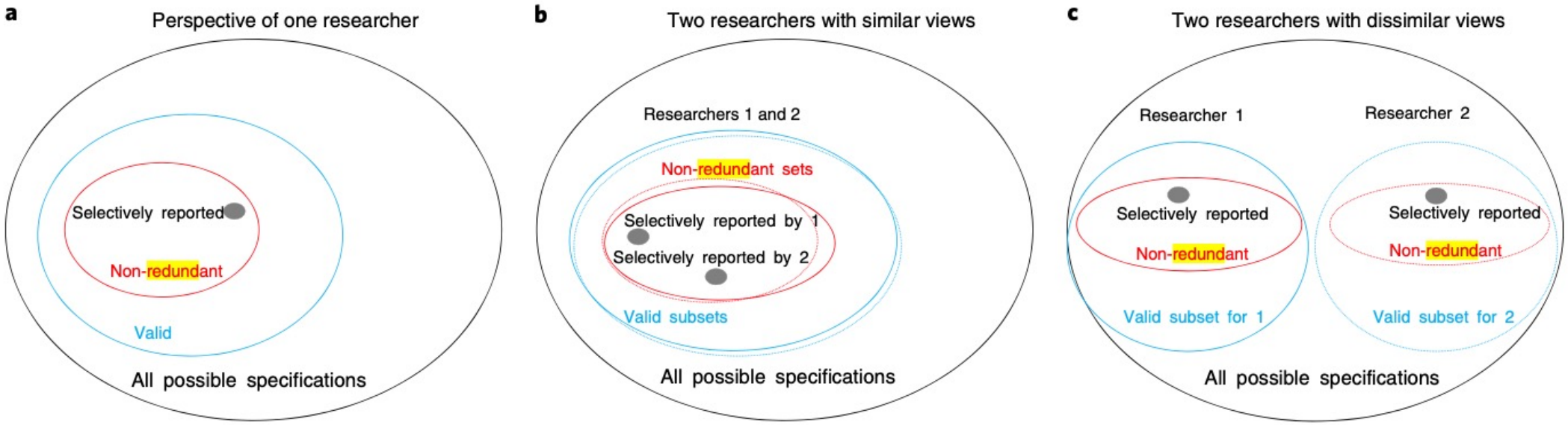
# But surely…

…you are not saying that you should run literally every possible configuration? Some of them will be stupid (e.g. omitting age in a study of offending) and some will be redundant (e.g. log-transforming x using either log(x + 1) or log(x + 1.1)).

Obviously not. The point is to explore the range of defensible and non-redundant specifications…



# But surely…




Because competent researchers often disagree about whether a specification is an appropriate test of the hypothesis of interest and/or statistically valid for the data at hand (that is, because different researchers draw different ovals), specification curve analysis will not end debates about what specifications should be run: specification curve analysis will instead facilitate those debates.

Even if two researchers have non-overlapping sets of reasonable specifications, specification curve analysis can help them understand why they may have reached different conclusions, by disentangling whether those different conclusions are driven by different beliefs about which specifications are valid, or whether they are driven by arbitrary selectively reported results from those sets. In other words, specification curve disentangles whether the different conclusions originate in differences regarding which sets of analyses are deemed reasonable (different red ovals), or merely in which few analyses the researchers reported (different grey dots).

# Thoughts?

Express them.