A Close Look at the Specification Curve Analysis and Existing Applications

---

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Wenxin Du

May 2020

Approved for the Division
(Mathematics)

Andrew Bray

Acknowledgements

I would like to thank my advisor, Professor Andrew P. Bray for the support and guidance throughout the thesis year, and for being patient, understanding, and supportive at this difficult time of the pandemic.
I would like to extend the appreciation to Professor Kelly McConville for the lively lectures and guidance on statistical projects and research. The skills I adapted from Statistical Practicum have been essential for the completion of this thesis.
I would like to also thank Professor Adam Groce, for being supportive and understanding during a stressful time.
I would like to thank every faculty, friend, and staff I have met during the four years at Reed. Especially my partner, thanks for all the companionship and support.

Preface

This thesis focuses on the Specification Curve Analysis, a method that assesses the robustness of scientific research models in response to changes in specifications. We address and evaluate the problems of one of its existing applications in Psychology. We also attempt to propose ideas and directions for further improvement of the SCA.

Table of Contents

List of Tables

List of Figures

Introduction

There is an awareness among scientists and researchers that the results of a scientific study can be greatly influenced by the "researchers' degrees of freedom", the arbitrary decisions made by researchers along the way of conducting data analysis. When conducting a study, researchers get to make decisions like choosing variables, method selection, etc. Considering all potential decisions as a set, then the decisions made by the researchers in an actual study may only be a small subset of it. However, every decision made during a scientific study can affect the study results, and the variability in making decisions would result in variability in study results. Many of the reproducibility problems existing in the scientific research fields are caused by such variability, as the significant results hinge on the specific set of decisions made. In these cases, the conclusiveness and generalizability of study findings can be limited. Realizing how researchers' degrees of freedom can affect scientific studies, methods have been proposed to work around such problems. This thesis will focus on a method called the Specification Curve Analysis (Simonsohn, Simmons, & Nelson, 2019), which attempts to assess the robustness of estimates in response to changes in operational decisions by identifying a set of reasonable decisions, gather the estimates and study the patterns among the estimates. There have now been several applications of SCA in the field of Psychology. With the hope of outputting reproducible and reliable results, psychologist Amy Orben conducted a study of the relationship between teenager mental well-being and digital technology use using SCA (A. K. Orben A. & Baukney-Przybylski, 2019). The study was published on Nature Human Behaviour, one of the top journals in the field, and was widely discussed and referenced. With a close look into the study, we notice several problems of its usage of SCA that may affect the interpretability and reliability of the study results. In this thesis, we attempt to replicate Orben's application of SCA and assess the problems existing in the application. Inspired by this work, we also attempt to propose ideas on further improvements of SCA.

Chapter 1

SCA and Its Applications

During the process of conducting a scientific study, researchers own high degrees of freedom on making data analytic decisions. They get to decide on things like the variables to be used, the outliers to be removed, etc. When there is an abundance of reasonable decisions to be made, the small subset of decisions chosen by the researcher can limit the conclusiveness of the study results. As an attempt to mitigate this effect of researchers' degrees of freedom on scientific study, Simonsohn et al. proposed the Specification-Curve Analysis (Simonsohn et al., 2019). The method considers the full set of "reasonable operational decisions" that can be made by a researcher, and reports result with an evaluation of its robustness in response to changes in choosing subsets of operational decisions. In this chapter, we focus on the Specification-Curve Analysis and its existing applications. We discuss the details of conducting an appropriate SCA and then introduce two existing applications of SCA.

1.1 Specification-Curve Analysis

For the following sections, we will follow closely the description of specification-curve analysis as proposed by Simonsohn et al. (Simonsohn et al., 2019). By Simonsohn, conducting a specification-curve analysis involves three steps: (1) Identifying the set of specifications, (2) Estimate over all the reasonable combinations of specifications and construct a descriptive specification curve, and (3) Conduct inferential analysis on a specification curve. In the following sections, we will discuss the details in each step, along with the important assumptions and concepts of the method.

1.1.1 Understanding and Choosing Specifications

The first step of conducting a Specification-Curve Analysis is to enumerate the set of *Specifications* to be considered. And before choosing the specifications, it's important to first understand what the term *specifications* mean, and the type of specifications

an SCA will be working with. *Specifications* usually refer to the decisions made by researchers while conducting a scientific study. Those may include deciding on a specific research question/statistical hypothesis, the choice of analysis method, operational decisions made during the modeling process, etc. The Specification-Curve Analysis requires a specific set of specifications which are: (1) consistent with the underlying theory, (2) expected to be statistically valid, (3) and not redundant with other specifications in the set.

It is required that the specifications used in an SCA are valid and non-redundant as determined by the researchers working on the study. Commonly, different researchers may consider different specifications as appropriate. When conducting an SCA, the researchers need only to consider the valid specifications from their perspective. If there is a substantial overlap between the valid specifications identified by different researchers, the results of the two SCAs will be similar. If the two sets minimally overlap, the results of two SCAs would expectedly be very different. As long as SCAs are applied appropriately, such a difference is likely not due to chance but may imply something fundamentally different between the two underlying theories.

One important concept about the Specification-Curve Analysis is that the method only works with specifications that are *operational decisions*, the decisions that do not affect the underlying theory but may affect the outcomes of the result. Say we are conducting an SCA studying the relationship between Y and X. SCA can work with specifications such as, "Do a log transformation on variable X", "Exclude three outliers", "Include variable K as control variable", "Add an interaction term between X and K", or "Do a logit model instead of a probit model". Such decisions do not change the statistical hypothesis or research question proposed beforehand. Instead, they can change model outputs and potentially lead to different analysis results. In other words, these specifications all focus on the type of operations that do not change the main characters and background in the story but may make small differences that can lead to a different story ending.

SCA does not work with specifications that are based on different underlying theories. For example, say we want to study the relationship between class performance and hair color, where the hair color refers to the natural hair color that is determined by genes. Using a variable that also considers dyed hair color would not be appropriate since the action of dyeing hair and the choices of colors can reveal information regarding personalities. The relationship between class performance and this variable can be different than the story we want to tell. Thus, the usage of variable "hair color appearance" as the independent variable will be an inappropriate specification to use for conducting an SCA on this research question.

A visualization may help with understanding this idea of operational decisions. Figure 1.1 presents as a tree the specifications that can be made by researchers when conducting a scientific study. In general, the researchers start by identifying a general area of interest (1) and then look for general topics that may be studied in the area (2). It is then possible to propose specific research questions, or statistical hypotheses in some cases (3). Once a specific question of interest is determined, an experiment may

be conducted to collect data, or existing datasets may be used for later analysis (4). With data collected, researchers may make a set of operational decisions on data and model (5). After all the steps are finished, the researchers collect the model outputs and can move to an analysis of the results.

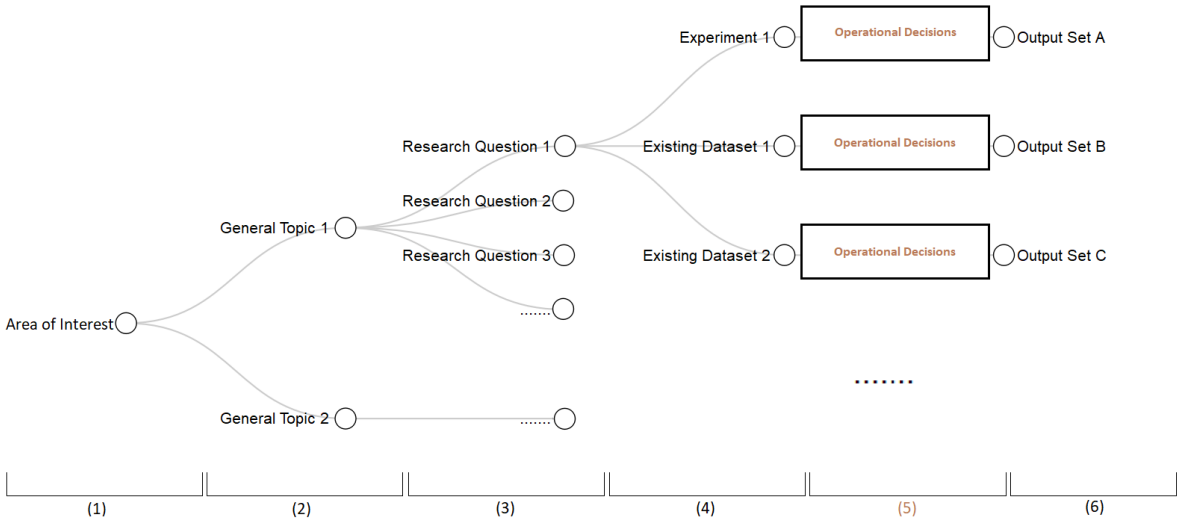

Figure 1.1: Visualization of scientific study steps

Each node on the tree represents a distinct decision made by the researcher. Each leaf of the tree represents essentially a unique set of research outcomes that can be produced by a specific set of decisions made along the way. When conducting an SCA, only the operational decisions inside one of the boxes are valid, and only one set of the outputs based on the same underlying theory and modeling is analyzed.

For example, a psychologist may be interested in studying the relationship between personal appearance and well-being. This would be a general area of interest. To conduct a study, the psychologist may then come up with several general topics, such as the relationship between personal appearance and mental health or the relationship between personal appearance and physical health. After careful consideration, the psychologist decides to focus on the first topic proposed. Within this general topic, multiple specific research questions can be proposed, which may include: “What is the relationship between hair color and teenager mental health”, “What is the relationship between piercing and mental health”, etc. After examining the existing literature, the psychologist decides to study specifically the relationship between hair color and teen mental health. Among the different ways of collecting data, the psychologist decides to conduct an observational study on hair color vs. teenager mental health. The psychologist then collects data and works on modeling and analysis. During this process, the psychologist comes up with a set of reasonable operational decisions based on their expertise in the field. Different combinations of the operational decisions will produce models that differ in some way but still are reasonable valid models. For example, say the field generally agrees on using model A to analyze a type of data,

but the psychologist has found novel literature suggesting using model B instead. Whichever model form chosen by the psychologist would be considered reasonable, but the choice may result in differed estimates. There would then exist a set of potential outputs that could be generated by choosing different combinations of the operational decisions. A visualization of this process is shown below:

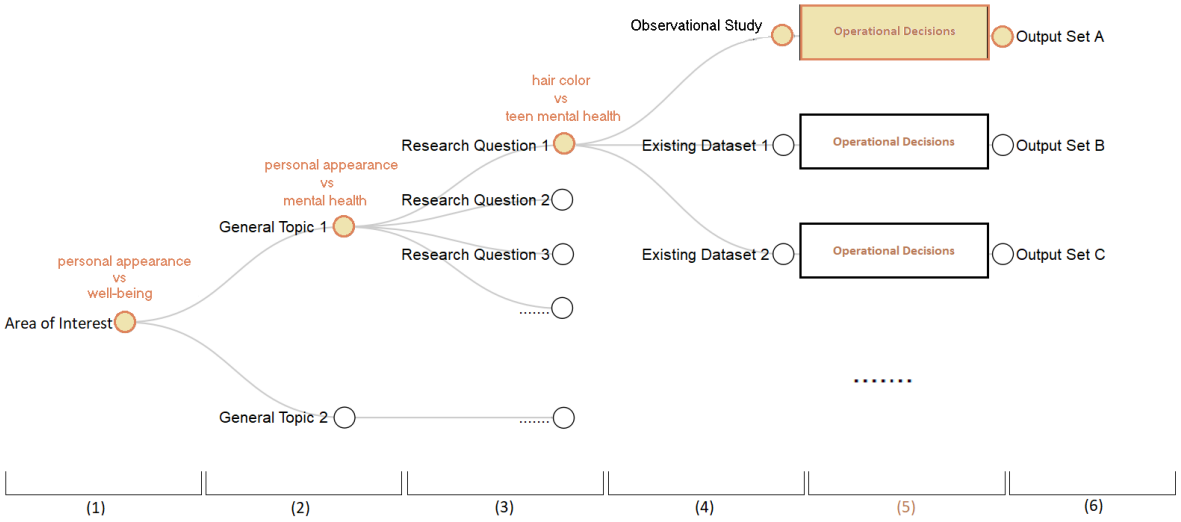

Figure 1.2: Procedure of a Scientific Study

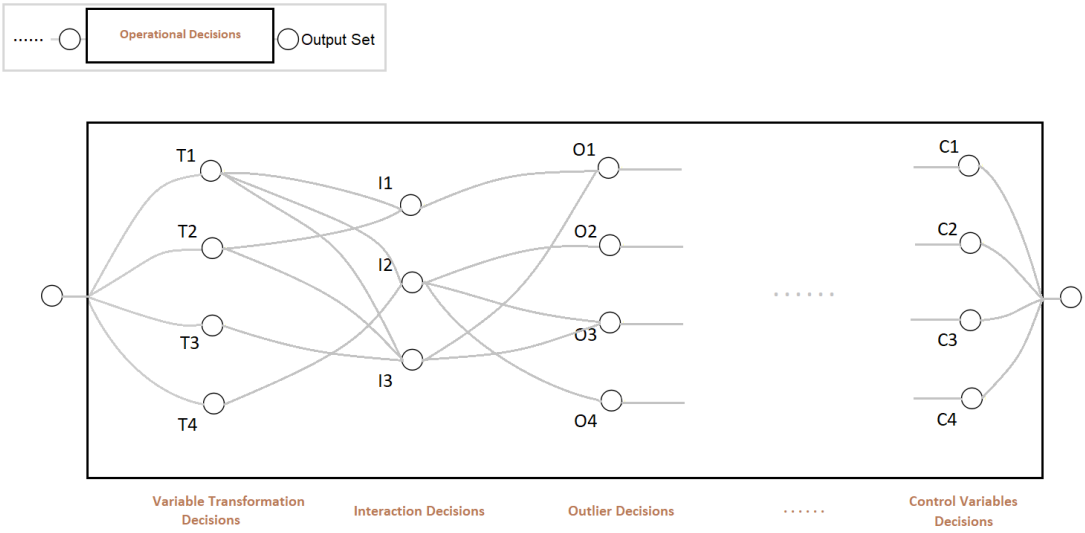

Figure 1.3: Visualization of Operational Decisions

Commonly in practice, the psychologist chooses one certain specification and produces the estimate from there. An SCA would consider all reasonable specifications and the full set of reasonable outputs. Figure 1.3 provides a closer visualization on the

operational decisions. In this figure, each node represents a unique decision that could be made by the researchers, and the branches connect appropriate combinations of decisions. Following the branches in different ways leads to different combinations of these operational decisions. In common practice, a researcher may choose one specific route on the tree that form up one specification, and consider the one outcome generated. An SCA attempts to consider all routes in the tree and consider the full set of outcomes.

Note that in figure 1.3, not all nodes are connected to nodes in the next group. In real life, not all combinations of operational decisions are appropriate to be applied together. For example, if a log transformation on a variable is performed, some data points may no longer be considered outliers and thus not removed. This variable transformation decision will not be used in combination with some of the outlier decisions and at least two nodes will not be connected by any branches. Ideally, the SCA will be working with only such appropriate combinations of operational decisions.

1.1.2 Constructing a Specification Curve

The next step is to build a specification curve. After determining the operational decisions, a set of specifications can be computed, where each specification leads to a different model to be run. Here we use a simple example to illustrate how to generate a set of specifications. Say a group of researchers considered only a set of model type decisions and a set of outlier decisions as 1) Use regression model A, 2) Use regression model B instead of A, 3) Use variable X as the independent variable, 4) Remove outliers from X and use the new variable X' as the independent variable. There will be four combinations of the two types of decisions and will produce specifications as:

1. Model A with independent variable X
2. Model A with independent variable X'
3. Model B with independent variable X
4. Model B with independent variable X'

When there are lots of variables involved, the list of specifications can be large, which can result in a huge number of total specifications. This can pose a real practical problem in terms of computation. For example, say we are working on a dataset with 10 variables, and say we identified: 1) 2 model decisions, 2) 20 variable transformation decisions, 3) 10 outlier decisions, and 4) 10 interaction decisions, this will result in 4000 different models. Running all 4000 models can take a while and can be computationally expensive with complicated model forms. It is also not rare for the number of variables to be much larger and the model form to be more complicated in real life. When a large number of specification models bring computational obstacles, a random subset of the specifications can be used instead.

Now that all the specifications have been determined, the next step is to tun all of the models and extract the point estimates. In the case of linear regressions, the extracted point estimates are generally $\hat{\beta}$ of each model. The estimates are then plotted as a curve, where the vertical axis refers to their numerical values, and the horizontal axis refers to the specifications that generated the specific model for this estimate.

As shown in Figure 1.4, a descriptive specification curve encompasses two parts: the top plot of a curve, and the bottom plot with lines and dots on it. In the top plot, the curve shows the estimates from each of the models, ordered from the lowest to the highest. Each model is produced by a specific set of specifications. The vertical axis represents the numerical values of the estimates, and the horizontal axis represents the set of decisions that produced the model, represented by dots. In the bottom half of the plot, each dot represents a operational decision. The vertical axis is the name of the decisions. For example, the first dot on the curve is an estimate from a model based on the decisions: "No controlling for year", "Main effect and no interaction", "Log Damage instead of Damage (in $)", "log-linear model instead of negative binomial model", "Use 0/1 for feminity instead of a 1-10 scaling", "Drop three hurricanes with highest damages as outliers", and "Drop two hurricanes with highest deaths as outliers".

Overall, it is possible to visualize from a specification curve if there exists a certain pattern relating to the choice of specifications and the corresponding estimation. For example, in the plot shown above, negative point estimates appear to require an idiosyncratic set of specifications. Also included in the plot is the indication of the models with statistically significant estimation. From the plot, it is possible to visualize if the statistical significance appears to be happening by chance, or if there appears to be some real relationship. For example, in this case, of the 1728 specification models, only 37 obtained statistically significant estimates. Overall, this specification curve may be suggesting that a non-statistically significant result is robust under alternative specifications.

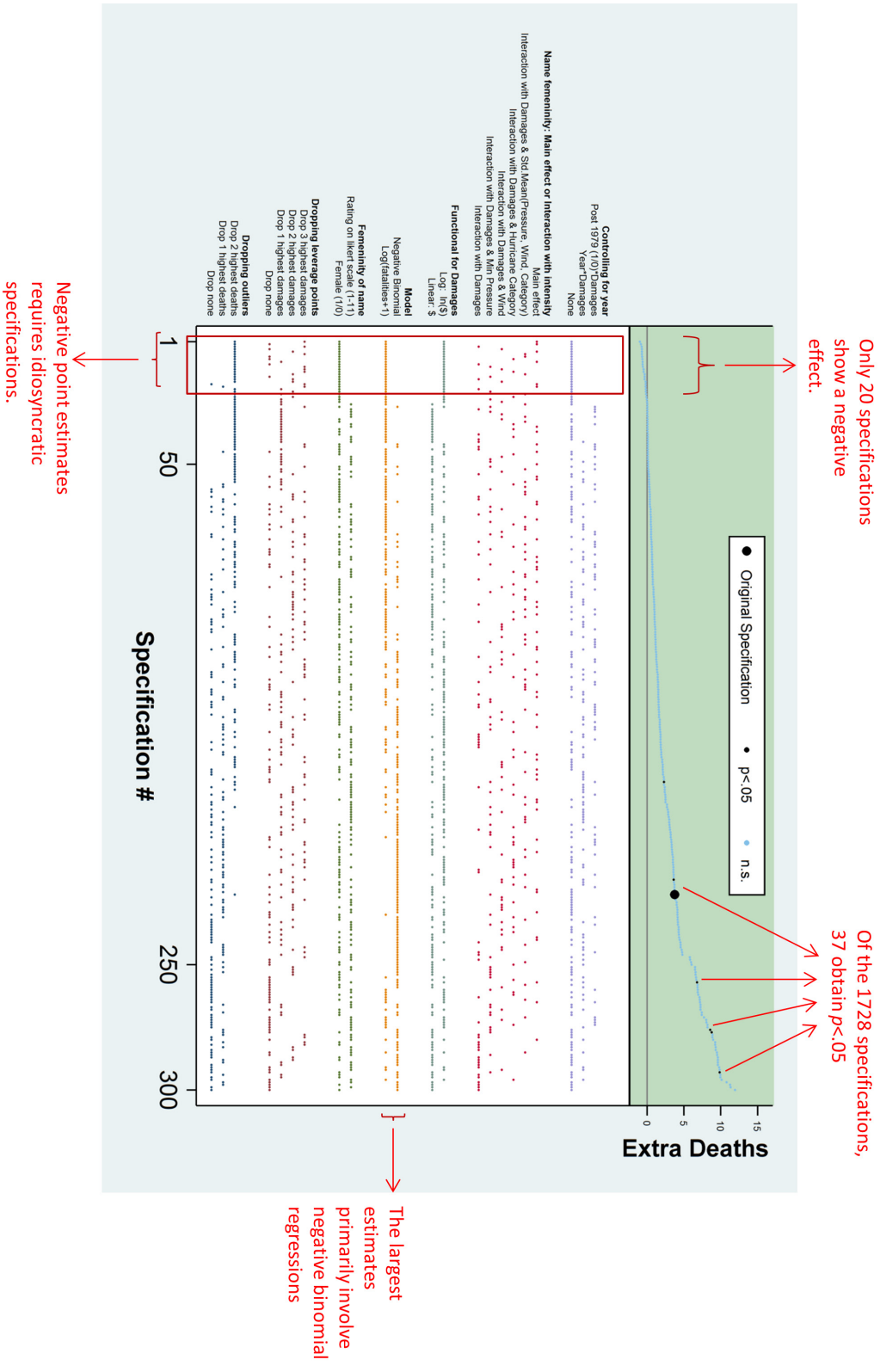

Figure 1.4: Specification Curve

1.1.3 Analysis on Specification Curve

The last step of an SCA is the statistical inference on the specification curve. The question for an inferential analysis was stated as "*Considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?*" (Simonsohn et al., 2019) The authors suggested that while the question is hard to be answered analytically, one can generate an empirical distribution of the specification curves under the null hypothesis using the technique of resampling. The examples provided in the paper all used the permutation technique for resampling of the data, and it was suggested that a bootstrapping technique can be applied for studies without random assignments.

Once the distribution of specification curves is obtained, three test statistics are proposed to do the inferential analysis: 1) the median overall point estimate from the specification curve, 2) the share of estimates in specification curve that are of the dominant sign, 3) the share that is of the dominant sign and also statistically significant ($p < 0.05$). The dominant sign here refers to the sign of the majority of estimates. If the majority of the estimates in an SCA have a positive sign, then the dominant sign will be positive. In case when there exists a non-zero relationship or effect, generally, we would not expect half the estimates to be positive and the rest to be negative, as the different models are not fundamentally different but rather similar at most places. The test statistic serves as a summary statistic of the entire specification curve.

Another tool for the inferential analysis is called a "confidence interval" of the specification curve. The confidence interval of the specification curve would be formed by two curves, call them the upper bound curve and the lower bound curve. To compute the curves, we take all sets of point estimates from the resampled datasets. For each set of point estimate, we sort them in order. The ith lowest estimate in the lower bound curve would be the $\frac{\alpha}{2}$ quantile of the ith lowest estimates from the sets of the estimates. For example, the minimum data point on the lower bound curve is obtained by first determining all the minimum estimates from the resampled datasets, and find the $\frac{\alpha}{2}$ quantile of these minimum estimates. A similar procedure will be taken to construct the upper bound curve, where the ith lowest estimate would be the $1 - \frac{\alpha}{2}$ quantile of the ith lowest estimates from each set of the estimates. It's suggested that in the case when the specification curve falls completely outside of the "confidence interval" curves, i.e. there is no overlap between the curves, the null is rejected.

Simonsohn et al. answered the research question using the test statistics in combination with the confidence interval curves. The followings are two examples given (Simonsohn et al., 2019):

Table 1.1: The test statistics results from two of the examples as provided by Simonsohn, where 2b) studied whether or not resumes with distinctively Black names lower callback rates when applying for jobs, and 2c) studied whether or not resumes with distinctively Black names benefitted less from higher quality.

Test Statistics	2b) P-Value	2c) P-Value
Median point estimate	< 0.002	0.030
share of estimates w/ dominant sign	0.125	0.130
share of significant estimates w/ dominant sign	< 0.002	0.162

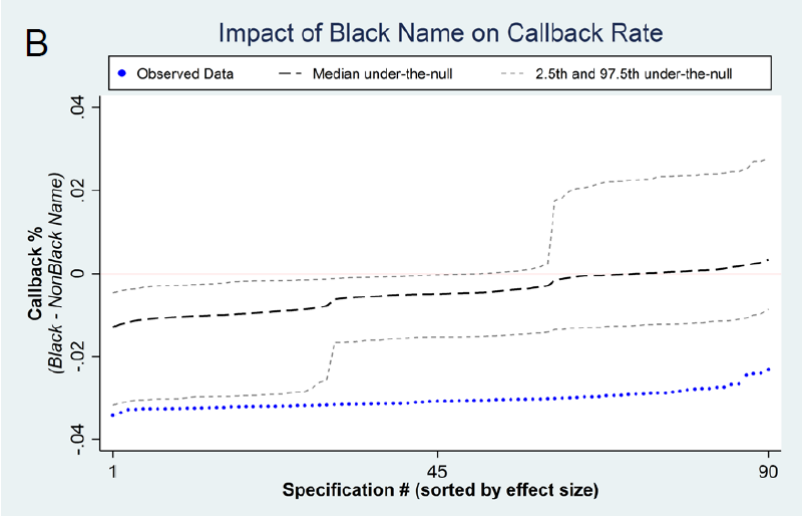

Figure 1.5: Confidence interval curves of the specification curves

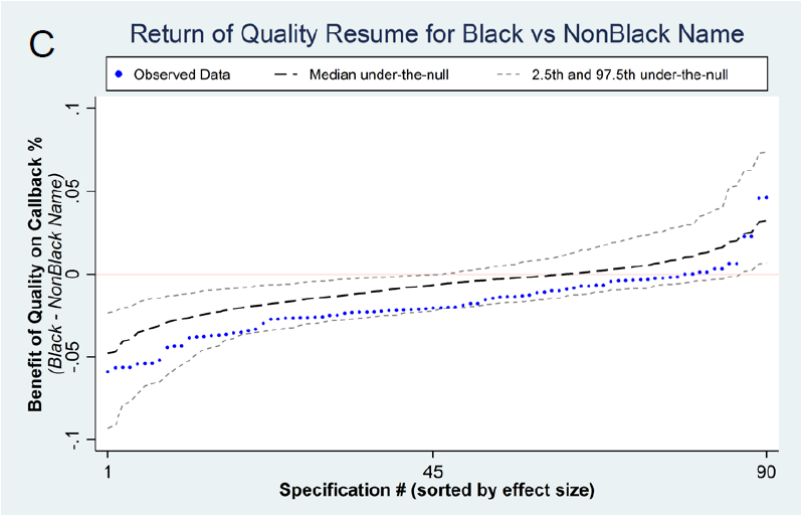

Figure 1.6: Confidence interval curves of the specification curves

Simonsohn et al. interpreted the results as the following:

> Starting with the core finding that distinctively Black names had lower callback rates we see that the entire observed specification curve falls outside the 95% confidence interval around the null. In the above table, we see that the null hypothesis is formally rejected.
>
> The robustness of the second finding, that resumes with distinctively Black names benefitted less from higher quality, is less clear. The observed specification curve never crosses the 95% confidence interval, and only one of the joint tests is significant at the 5% level.

In study 2b), two of the test statistics produced statistically significant p-value, and the specification curve falls completely outside of the confidence interval curves. The authors thus rejected the null hypothesis. In the other case, the curve overlaps with the confidence interval curves and there exists no statistically significant p-value among the three test statistics. The authors claimed that the robustness of the existence of a relationship is less clear in this case, which we interpret as no sufficient evidence to reject the null hypothesis.

Simonsohn et al. did not specify whether or not the two inferential methods must be used in combination or can be used separately. Moreover, the confidence interval curves method was only mentioned in a caption of the example plot of the confidence interval curves. No discussion on the rationality of this inferential method was provided. The curves are not specification curves, thus the curves are in fact not bounds of a "confidence interval" of a specification curve. The ith lowest data points on the confidence interval curves can be understood as the empirical confidence interval of the ith order statistics of the estimates from the specification curve. More further studies will be needed to assess the rationality of the usage of this method. The existing applications of the SCA, in which we will talk in more detail in the next section, have not adapted the confidence interval curves inferential method. In this thesis, we will be focusing on formalizing and applying only the inferential hypothesis test, and leave the "confidence interval curves" method for future study.

To understand better how SCA has been applied in real life, in the following sections, we introduce the existing applications in the field of Psychology with an emphasis on two applications. We also provide a more detailed discussion of the applications in the next chapter.

1.2 Applications of SCA

The paper that proposed SCA is published as a working manuscript in 2015, and the method has already been widely applied in the field of Psychology. A few published studies have used SCA to study topics including the effect of social media on adolescent life satisfaction (A. Orben, Dienlin, & Przybylski, 2019), the relationship between

adolescent mental health and technology use (A. K. Orben A. & Baukney-Przybylski, 2019), the association between digital-screen engagement & adolescent well-being (A. Orben & Przybylski, 2019), the effect of birth-order position on personality (Rohrer, Egloff, & Schmukle, 2017), etc. In this section, we focus on two of the applications and discuss the design of SCA analyses in the studies.

1.2.1 Adolescent Mental Health and Technology Use

This study conducted by Orben and Przybylski attempt to assess the association between digital technology use and adolescent well-being using 3 large-scale social datasets: Monitoring the Future (MTF) (Johnston, Bachman, O'Malley, Schulenberg, & Miech, 2017), Youth Risk and Behaviour Survey (YRBS) (Kann L, 2015), and Millennium Cohort Study (MCS) (University of London, n.d.). The data were collected from studies of the same names and encompass survey answers from adolescents and relatives on a variety of topics over a long period. For each of the three datasets, Orben identified a set of specifications and conducted SCA analysis for the research question of "the association between adolescent well-being and digital technology use". In this section, we summarize Orben's approach, main steps, and main findings. A detailed assessment and critique of the usage of SCA in this study will be provided in Chapter 2.

Identifying Specifications The first step to conduct an SCA analysis is to identify the set of reasonable specifications. The following picture is a screenshot from Orben's study about the specifications identified:

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

Figure 1.7: Specifications identified in Orben's study

As we can see, in this study, there are mainly three types of specifications considered: 1. variables representing adolescent mental well-being; 2. variables representing digital

technology use by individuals; 3. whether or not to include a set of predetermined control variables. Two specifications are not of the above categories: "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" for YRBS and "Use well-being measures declared by cohort members or those declared by their caregivers" for MCS. The former is a specification on the encoding of the dependent variable. And the latter is the specification of choosing between the survey answers about the mental well-being of teenagers and those from parents.

The model is set in default to be linear regression. When the control variables are in use, the model will be multivariate linear regression. Otherwise, it will be just a simple linear regression. It is worth noting that, when choosing the set of variables representing digital technology use, Orben considers multiple variables on different types of technology as alternatives to each other. As shown, the list of alternative variables to represent "digital technology use" for MCS includes: "Whether or not own a computer at home", "Hours of social media use on a normal weekday", "time on TV viewing on a weekday", etc. One model for MCS may use "Whether or not own a computer at home" as the independent variable, while the other model may use "time on TV viewing on a weekday" as the independent variable. Similar choices of alternative variables to represent "digital technology use" are made for the other two datasets. A total number of 372 specification models were determined for YRBS, 40,966 specification models were determined for MTF, and 603,979,752 specification models were determined for MCS. In the case of MCS, and a random subset of the specifications models with size 20,004 was used instead for computational purposes.

Single SCA and analysis After determining the set of specifications for each of the three datasets, three specification curves were generated. For each fitted specification model, the estimate of β on the variable representing "technology use" was collected and presented on the curve. Instead of focusing on the curves, Orben analyzed the summarized statistics from the specification curves. She focused on the sign and magnitude of the median β estimates and concluded that a small negative relationship is determined for each of the three datasets. A full table of results will be included.

Bootstrapping test and inference The last step is to conduct inference on the SCA result. Orben performed a bootstrapping test and generated 500 specification curves on bootstrapped data. The inference was performed using all three test statistics as suggested by Simonsohn et al. The p-values found were all approximately 0. As a result of the test, she concluded that evidence has been found supporting the negative relationship between digital technology use and adolescent well-being. The method of confidence interval curves was not used in this application.

Others In addition to the SCA analysis on the research question, Orben performed additional SCA analyses on the relationship between adolescent mental well-being and several other variables of interest, including binge-drinking, smoking marijuana, being bullied, arrested, perceived weight, eating potatoes, etc.. The mean estimates on technology use variables is compared with these results, and it is suggested that

the small negative effect of technology use on adolescent mental well-being may be too small to warrant policy changes.

However, several major issues exist in the application of SCA in this study calling into question the reliability of the results. In Chapter 2, a description of the full replication of the study along with detailed assessments and critiques of this study will be provided.

1.2.2 Birth-Order Position and Personality

We now introduce another application of the SCA in the Psychology field. The application of SCA follows more closely the steps introduced by Simonsohn et al., and we will be using it as a comparison to Orben's.

Rohrer et al. applied SCA and studied the effect of birth-order position on personality. The data used in this study came from the SOEP, which is an ongoing study of private households in Germany and their members. The study focuses on multiple research questions of interest, studying the effect of birth-order position on 11 personality variables, including life satisfaction, interpersonal trust, intellect, etc. In comparison to the description of Orben's study, we will focus less on the details but instead focus on its main steps of constructing the SCA's. A comparison between this application and Orben's application will be provided in the next chapter.

Identifying Specifications An SCA was run for each of the 11 research questions. A different set of specifications was determined for each of the SCA. The paper provided a list of the model specifications determined:

1. Different ways to measure the personality variable;
2. Use raw scores or age-adjusted scores;
3. Within-family or between-family analyses;
4. Which definition of birth-order position to use: the social definition or the more restrictive definition limited to full siblings;
5. Differentiation of each birth-order position within a sibship (e.g., first, second, third) or differentiation only of firstborn from later positions;
6. Inclusion of all sibships or sibships with spacing does not exceed 5 years, or sibships with sibling spacing exceeded 1.5 years but did not exceed 5 years between any two siblings;
7. Exclusion of any gender effects, the inclusion of the main effect of gender, or inclusion or both the main effect of gender and the interaction of birth-order position and gender;
8. Analysis of the complete sample, analysis of only individuals from sibships with 2 to 4 children, or separate analyses for sibships of 2, 3, and 4 children.

Many of the specifications considered in this study are appropriate operational decisions, including outlier decisions and variable transformation decisions. It appears that some specifications identified may be based on the different underlying theories and/or different research questions. In the next chapter, we will discuss in more detail the choice of specifications in this study.

SCA and analysis For most of the 11 SCAs, 720 specification models were determined, while two of them determined a larger number of specification models: 1440 and 2160. All models were run and the estimates of the main effect were extracted for analysis. A permutation test is performed for inference, following the same procedure as performed in the examples provided by Simonsohn et al. All three suggested test statistics were used. The p-values are then used for evidence of a statistically significant effect. The confidence interval curves method was not used in this application.

There are several distinctions between the applications of SCA in these two studies. In the next chapter, we will look closely at each step of the two applications and compare the applications to the procedure proposed by Simonsohn et al.

Chapter 2

Replication and Evaluation

In this chapter, we discuss the attempt to replicate Orben's study along with the assessment of its use of SCA. We also make comparisons of Orben's approaches with the work by Rohrer et al. in terms of applying SCA. We start from a more detailed introduction of the study and its aims, and then we discuss the details of using SCA in the study. We will introduce the three datasets used, the attempt to replicate Orben's process of applying SCA on the three datasets, and the reproduction of SCA results. We will include also an discussion on the obstacles to overcome during this process. After discussing the replication, we evaluate the problems existing in Orben's application of SCA.

2.1 Publication and ReproducibiliTea

Before getting into the details in the application of SCA, we would like to introduce some details about the publication of this study and the great efforts made by the authors to work in a transparent manner and encourage other researchers to engage with their work. Without these efforts, this replication would not be possible.

The study *The association between adolescent well-being and digital technology use* (A. K. Orben A. & Baukney-Przybylski, 2019) was published in 2019 in the journal Nature Human Behaviour. It is an online journal that publishes "research of outstanding significance into any aspect of individual or collective human behavior". The journal covers topics from Social Science, Neuroscience, Health Science to Physical Science, and is among one of the top influential journals in these fields. Research studies being published by the journal are known for having high quality and being related to the most pressing social problems and topics. Articles and research published in the journal often are those having outstanding findings and influence related fields. Orben's study, published in January 2019, has already had 152 citations (as of April 2020).

Tremendous efforts have been made by Dr. Orben to enable the reproduction of her

analysis. The datasets used for this study can all be obtained through public sources. The data wrangling process, along with all code used to produce the SCA results can be found on a public GitHub repository (A. Orben, 2019a). Orben used R for all the coding, and detailed comments were provided on all coding files. The efforts have made a replication and reproduction of the study easy and straightforward. This thesis would not have been gone as smoothly without these efforts.

The study was conducted with the intention of reducing the impact of researcher degrees of freedom and producing more reliable and robust scientific study results. Aiming at reproducibility and open science, Dr. Orben, along with two other scientists, started a journal club ReproducibiliTea in early 2018 at the University of Oxford.

> We hoped to promote a stronger open-science community and more prominent conversations about reproducibility. The initiative soon spread and is now active at more than 27 universities in 8 countries. – Dr. Orben (A. Orben, 2019b)

The solid aim for openness and reproducibility shown by Dr. Orben greatly encouraged the completion of this thesis, we appreciate and value the efforts. As indicated in the last chapter, we have noticed problems in this application of SCA which challenge the reliability of its results. By indicating the existence of these problems, we wish to provide advice from a different perspective and hope to provide help in producing more robust and reliable results. In the following sections, we describe the replication process and obstacles and provide a detailed look and assessment of this application of SCA.

2.2 Data and Reprocessing

Three large-scale social datasets were used in Orben's study: Millennium Cohort Study (MCS) (University of London, n.d.) from the United Kingdom, Youth Risk and Behavior Survey (YRBS) (Kann L, 2015) from the US, and Monitoring the Future (MTF) (Johnston et al., 2017) from the US. The three datasets were all survey data obtained from the scientific study of the same name, and encompass survey answers from adolescents aged predominately 12-18 from 2007 to 2016. The datasets provided wide measures of adolescents' psychological well-being and digital technology use. A considerable number of psychology studies in the existing literature were conducted based on large-scale studies, which provided a wide selection of approaches to modeling and analysis based on the specific dataset. In this section, we discuss the background information of the three datasets and the reprocessing of the data obtained from public sources.

2.2.1 YRBS

The Youth Risk Behavior Surveillance YRBS was first launched in 1990 as a biennial survey of adolescents that reflects a nationally representative sample of students attending secondary schools. Orben's study focused on data collected from 2007 to 2015, which we were able to obtain from the YRBS website (Kann L, 2015). While Orben used data in SPSS format, we were only able to access the data through Microsoft Access. The datasets were extracted and saved under excel format. It was confirmed that the same number of observations were included in the obtained dataset as the data used by Orben, 37,402 girls and 37,412 boys from 2007 to 2015. It was also confirmed that all variables used in Orben's study are contained in the obtained dataset. Most of the work in the preprocessing step for YRBS focused on transforming the data types from character strings to numerical values.

One noticeable obstacle in this step was that, since the study is conducted annually and is still ongoing, the survey questions and indexings have been updated several times in recent years. The majority of the variables in the datasets are named after the survey questions indices and the recent updates in survey questions result in different indices for survey questions between the current survey and surveys conducted before 2015. This leads to mismatches between variable names in the incorporated dataset including data from the year of 2015 and prior (the one used by Orben) and the variable names in the dataset obtained for this study, including data from the year of 2017 and prior. Only the data from 2015 and prior were used for this replication. Careful research and recoding are done to ensure the correct set of variables was used for the replication.

2.2.2 MTF

Monitoring the Future was first launched in the year of 1975 as an annual nationally representative survey of approximately 50,000 US adolescents in grades 8, 10, and 12. The data are publicly accessible through ICPSR, the Inter-university Consortium for Political and Social Research (Johnston et al., 2017). Surveys on adolescents in grade 12 were not used in the analysis since "many of the key items of interest cannot be correlated in their survey". (A. K. Orben A. & Baukney-Przybylski, 2019) Orben focused on the data collected from 2008 to 2016, which included 136,190 girls and 132,482 boys. While the MTF data for each year is publicly accessible, no merged MTF dataset for the specified period was found in the authors' repository. From 2008 to 2016, the survey was updated multiple times, along with one major change in data file format after RStudio's release in the year of 2011. Before then, the data files were available in formats including SAS, SPSS, Stata, etc, but were only available as separated sub data files. After Rstudio, the data can be obtained from one complete R data file for implementation to R. Due to the frequent updates in the annual surveys and changes in data files, the variable names vary greatly among the available datasets. This made it excessively difficult to recreate the same dataset used by Orben. After

getting into contact with Dr. Orben, luckily, a merged data file was obtained. However, due to time restraint, the replication on MTF datasets was not completed by the time of completing this thesis, and thus not included.

2.2.3 MCS

The Millennium Cohort Study follows a specific cohort of children born between September 2000 and January 2001 and collects data from both the children and the caregivers. Orben's study focused specifically on the data collected in 2015 when the children were between ages 13 and 15. The sample included 5926 girls and 5946 boys along with 10605 caregivers. We were able to obtain this same dataset through CLS, the Centre for Longitudinal Studies at UCL Institute of Education (University of London, n.d.). Access to the data is open to registered users of UK Data Service with the agreement to the terms of use. While Orben obtained data in CSV format, we were only able to obtain data in SPSS format. The same set of observations, with 5926 girls and 5946 boys born between September 2000 and January 2001, were included in the dataset, along with the same set of variables as used in Orben's study.

Unlike working with YRBS and MTF, the variable names in the obtained dataset matches well with the variable names in the dataset used by Orben. However, instead of using numerical indices to represent survey answers, in the dataset obtained, the variable values were all in characters. After careful reprocessing, all variable values were transformed and matched with the numerical values of the variables as were in Orben's study. However, two variables–one related to family incomes and one related to siblings–had only NA values in the obtained dataset. The omissions might be done for confidential purposes. The two variables were used as control variables in Orben's study. As we fail to obtain the two variables, they were removed from this replication.

2.3 Replication and Reproduction

The replication of Orben's analysis consists of two parts: the replication of generating a single specification curve for each dataset, and the replication of the inferential specification curve analysis, which assesses the significance of the single SCA result. The code used for Orben's study is publicly available on the Open Science Framework website (A. Orben & Przybylski, 2020) and her GitHub repository [OrbenRepo], and all replications were performed based on the provided code. In the following section, we discuss the procedure, obstacles, and specific resolutions to the obstacles of replicating the analysis.

2.3.1 Generating SCA curves

The first part of the replication is to replicate the single SCA analysis for each dataset. While all work done in this section is based on the code provided by Orben, due to the necessary reprocessings mentioned in the previous sections, slight modifications were made for smooth replication.

As mentioned in Chapter 1, three types of specifications were identified by Orben. Based on the public code, we were able to obtain the same set of specifications as used in Orben's study. A note-worthy obstacle is that, due to a large number of specification models determined for the MCS study, a random subset of 20,004 specification models was used instead. A seed is not provided by Orben for the random subset, thus we failed to obtain the same subset of specification models for this SCA analysis. We instead randomly generated our subset of 20,004 specifications. This randomness may result in a discrepancy in this specification curve. Considering that the random subset has a large size, we expect the degree of this discrepancy to be small. This expectation is confirmed by replication result: while Orben obtained the median coefficient of the independent variable to be $\text{Median}(\beta) = -0.032$, our replication obtained $\text{Median}(\beta) = -0.0328$.

The problem does not exist for the YRBS study. There were fewer variables available in the dataset relating to technology use and adolescent mental well-being. The number of specifications identified in the two studies is smaller, therefore the exact set of specifications was used for the replications. The result matched well with Orben's result. The median coefficient of the independent variable in the YRBS study was found to be $\text{Median}(\beta) = -0.035$ in Orben's study. The result obtained in this replication, when rounded to the same digits, is also -0.035.

2.3.2 Inferential Analysis

The next part of the replication is to replicate the inference of the specification curves for each dataset. Orben chose to use a bootstrapping test on the median overall point estimate for the significance of the result. We will later assess the choice of the inference test and the validity of the inference. For now, we focus only on replicating the test and the result.

As described in Chapter 1, a resampling technique is suggested to be used to generate a null distribution of specification curves for reference. Orben used the Bootstrapping technique and produced 500 bootstrapped samples for each dataset. As described, she chose the test statistics method to make inference and analysis. It was found that the three test statistics, including the median overall point estimate, the proportion of estimates with dominant sign, and the proportion of significant estimates with dominant sign, were all statistically significant. The initial attempt of the replication was done using the original code as provided on the OSF website. However, due to the large sizes of the three datasets and the great number of loops used in the R code, the

replication process was extremely computationally intense. A single specification curve will take around 8 hours to be generated on 1 core, and performing 500 specification curves will take nearly 24 weeks. An ARC computer cluster at Oxford was used by Orben to reduce running time, however, no access to such an advanced computer is available for this replication. Therefore, instead of using purely the original code, the code for this replication was rewritten to run in parallel. The running time has been significantly reduced. The dataset YRBS has the least number of observations and specifications, and after the recoding, it now takes about 9 hours to generate a complete bootstrapping distribution of 500 specification curves on a Rstudio server with 8 cores. More time will be needed for the other two datasets, as the number of observations and specifications can be much higher in those two cases, but still within a computationally reasonable time range.

As mentioned earlier, a seed was not provided in Orben's study. Therefore we cannot fully replicate the randomness of a bootstrapping test. The bootstrapped samples in this replication are different than the samples used in Orben's work, and this may result in a difference between the results of Orben's and this replication. In Orben's study, the resulted bootstrapping distribution of specification curves produced p-values of 0.00 for all three test statistics for all three studies. In our replication for YRBS and MCS, we obtained the p-values being less than 10^{-6}, which are approximately 0. This suggests the same result as the original inferential test result, where the test statistics are shown to be statistically significant and we can reject the null hypothesis of no effect, $\beta = 0$.

2.4 Evaluating Orben's work

A replication allows a full understanding of Orben's approach and procedure. It is only when we have a full understanding of the work that our critiques and assessments on it will be responsible and well-informed. In this section, we talk in detail about the existing problems in this application of the SCA method, including some fundamental misunderstanding of the intentions and applicabilities of the SCA method, inappropriate choice of specifications, and misinterpretation of the SCA results. We also compare Orben's procedures with the procedures taken by Rohrer et. al when studying the effect of birth-order position on personality, which is considered a more reasonable and appropriate application of the SCA method.

2.4.1 One Research Question or Many?

We start from assessing the research question of this study. The article is titled "The association between adolescent well-being and digital technology use". As addressed in the paper, the main focus of this paper is to study the association between digital technology use and adolescent well-being. This is a broad topic to be studied. Digital technology is a general category of many things, including interactive digital technology

such as social media platforms, and non-interactive digital technology such as TV. Is it reasonable to consider the different types of digital technologies as a unity and study its relationship with teenagers' mental well-being, or could the different types of digital technologies be having different relationships with teenagers' mental well-being?

It has been studied in the field of Psychology that categorizing certain types of digital technology use into a broader overarching category is inappropriate. Studies suggest, for example, that categorizing the different types of internet activities into an overarching category is suboptimal. Bessiere et. al found results suggesting that differences in social resources and choices of how people use the internet may account for different outcomes in measure of depression. (Bessiere, Kiesler, Kraut, & Boneva, 2008) (Bessiere, Kiesler, Kraut, & Boneva, 2008) Burke et. al looked more specifically on how using the internet passively or actively can result in different outcomes in social communication skills and self-esteem. (Burke, Kraut, & Marlow, 2011) (Burke, Kraut, & Marlow, 2011) Similar results were also obtained by Verduyn et al. (Verduyn, Ybarra, Resibois, Jonides & Kross, 2017) (Verduyn, Ybarra, Résibois, Jonides, & Kross, 2017). The existing literature suggests considering the active usage of media and passive usage of media as having a different effect on mental health and other personality scales. The different types of technologies allow a different level of engagement and interactions. Based on the suggestions from the literature, it may not be appropriate to consider the different types of technologies as representative and interchangeable to represent the general usage of technology.

Now let us think about Orben's choice of the set of alternative variables representing "digital technology use". The list of alternative variables to represent "digital technology use" for MCS includes: "Whether or not own a computer at home", "Hours of social media use on a normal weekday", "Time on TV viewing on a weekday", etc. "Whether or not own a computer at home" measures the availability of a computer to an individual, but not the way an individual uses the computer. But "Hours of social media use on a normal weekday" and "Time on TV viewing on a weekday" measures the active usage of using social media or watching TV on a weekday. Not all variables here actually measures "digital technology use". Even for those variables measuring usage of digital technology, based on the existing literature, the usage of different technologies may have a different effect on mental health. These models may be constructed based on different underlying theories.

Recall from Chapter 1, that the appropriate set of specifications considered for an SCA analysis is a set of operational decisions specific to a pre-determined research question and study design. Visualization of such a set of specifications was shown in figure 1.2. In this case, however, instead of conducting a study based on a specific research question, Orben may have considered multiple of them as alternatives to each other.

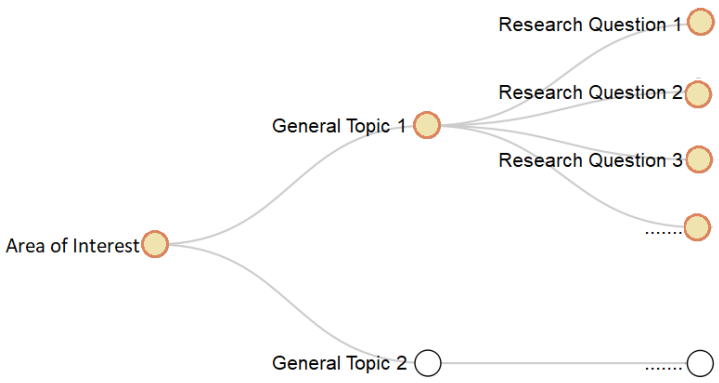

Figure 2.1: Orben considers multiple research questions as interchangeable with one another instead of choosing a specific research question from the general topic, contrary to the recommendation of Simonsohn et al.

Such a problem does not exist in the study of birth-order position and personality. (Rohrer et al., 2017) The researchers considered 11 specific research questions and conducted a separate SCA for each. The research questions pre-determined the specific variables of interest. For example, one of the research questions studies the effect of birth-order position on life satisfaction. Different scales may be used for measuring life satisfaction, but they are all reasonable measures of life satisfaction, a specific aspect of personality.

2.4.2 Choice of Specifications

The specifications determined by Orben, due to the consideration of multiple research questions, are indeed specifications in light of different underlying theories. While one specification suggests using the variable on TV used to represent general digital technology use, a different specification suggests using the variable on electronic games use to represent general digital technology use. The stories told by these different models generated by the different specifications may be very different. When performing SCA on such specifications, it's not only the impact of arbitrary operationalizations of the models that are moderated but also the impact of non-arbitrary theorizing that's moderated. This conflicts with the true intention and appropriate usage of SCA.

It's also worth mentioning that the specifications determined by Orben in this study are mainly specifications relating to the inclusion/exclusion of variables. The determined specifications can be mainly categorized into three types: 1) specifications on the choice of the dependent variable, 2) specifications on the choice of the independent variable,

3) whether or not to include a pre-determined list of control variables. However, the SCA should consider a full set of combinations of operationalization decisions instead of just those of variable selections. Important operationalization decisions, such as the recoding of the variables as performed by Orben in the data processing step before actual analysis, are decisions that can have an important effect on the result and are not being considered in this study.

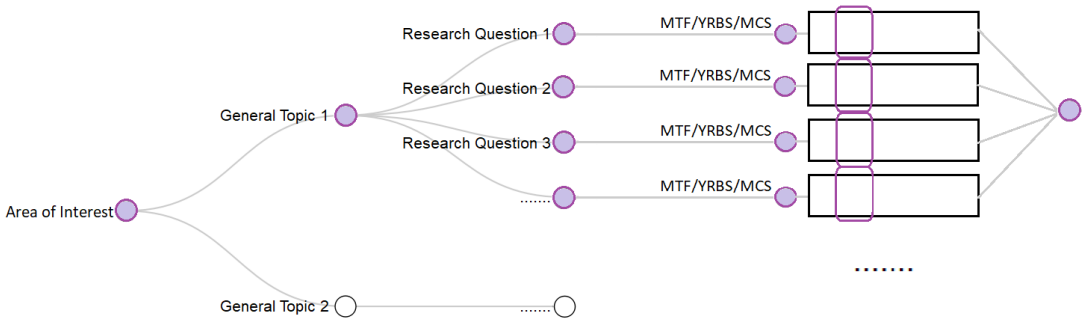

Figure 2.2: The set of specifications considered by Orben

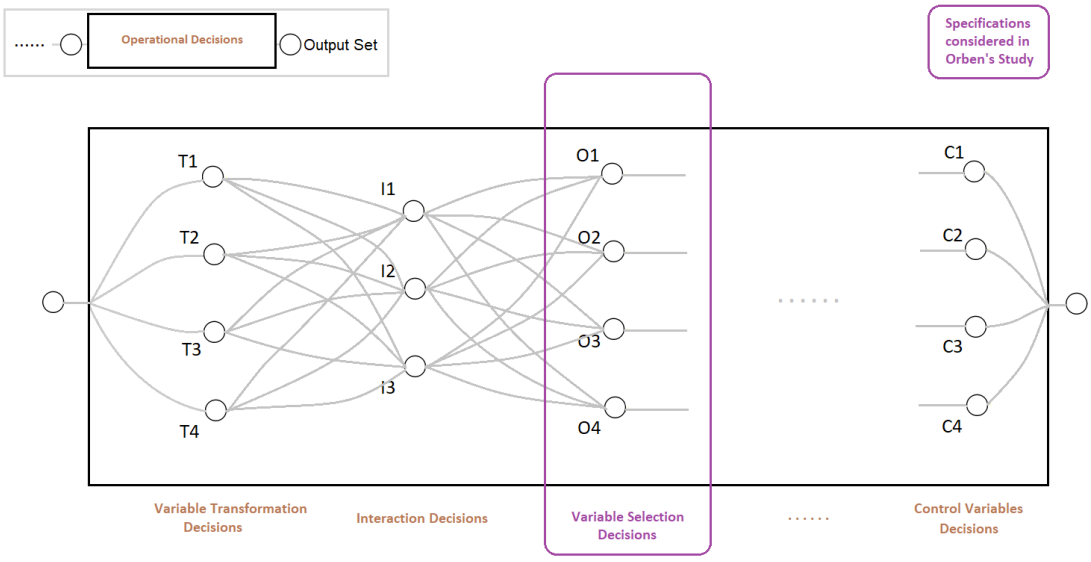

Figure 2.3: The set of specifications considered by Orben

The above figures provide an overall visualization of the specifications considered in Orben's study. Instead of considering one full set of combinations of operational decisions under a specific research question and study design, Orben considered a subset of operational decisions under multiple research questions. The outputs from

the different branches are collapsed together into one analysis. This is, very clearly, different from the usage recommended by Simonsohn et al.

We can compare this choice of specifications to the one by Rohrer et al. The following is the list of the specifications determined in the paper.

1. Different ways to measure the personality variable (the one out of 11);
2. Use raw scores or age-adjusted scores;
3. Within-family or between-family analyses;
4. Which definition of birth-order position to use: the social definition or the more restrictive definition limited to full siblings;
5. Differentiation of each birth-order position within a sibship (e.g., first, second, third) or differentiation only of firstborn from later positions;
6. Inclusion of all sibships or sibships with spacing does not exceed 5 years, or sibships with sibling spacing exceeded 1.5 years but did not exceed 5 years between any two siblings;
7. Exclusion of any gender effects, the inclusion of the main effect of gender, or inclusion or both the main effect of gender and the interaction of birth-order position and gender;
8. Analysis of the complete sample, analysis of only individuals from sibships with 2 to 4 children, or separate analyses for sibships of 2, 3, and 4 children.

The specifications determined above can be mainly categorized into the following types: variable measurement decisions, variable transformation decisions, and outlier decisions. Different combinations of specifications do not change the research question being studied and are mainly reasonable alternative ways of analyzing this specific topic.

2.4.3 SCA interpretation

The last major issue of this study is the way Orben interprets the SCA result. As discussed in the previous chapter, the specification curve provide information on whether or not there exists a robust relationship in response to changes in specifications. The specification curve is not used for interpretation of the actual magnitude of the numerical values of the estimates. However, in Orben's study, when interpreting the specification curve, the median values of the β estimates were used and the magnitudes of the numerical value were considered. Here is a quote from the study:

> The SCAs showed that there is a small negative association between technology use and well-being, ...

The "SCAs" here refers to the single specification curve generated for each of the three datasets, MTF, YRBS, and MCS. And the "small negative association" was

concluded from the median estimate of the β's from models with changing specifications. Simonsohn et al. suggests using the specification curve to look for evidence of a robust relationship in response to changes in specification, but does not mention interpreting the numerical values of the point estimates' summary statistics. The three examples provided in the original paper describing the method only used the specification curve to assess 1) if the relationship seems strong and 2) if an outstanding set of specifications producing similar results exists: 1) In the case when the majority of estimates are of one sign and are statistically significant, the authors consider it to be evidence for a robust relationship; 2) if a set of similar specifications tend to produce outstanding estimates, one may consider if it suggests the existence of an unnoticed underlying theory. When conducting inferential tests for the three examples, the medians were only used to check for statistical significance. The numerical values of the medians were never considered to be meaningful.

It is worth mentioning here that the application performed by Rohrer et al. did not interpret the numerical values of the test statistics. The analysis interpretations of the SCA results follow closely the examples provided in the work of Simonsohn et al. In Chapter 3, we formalize the SCA based on the procedure described by Simonsohn et al. We then discuss the different ideas for improvements and additional inference on SCA, which will include a detailed discussion of why the departures that Orben takes from Simonsohn et al. may provide unreliable results and interpretations.

Chapter 3

Formalization of SCA

In this chapter, we discuss an overall formalization of the Specification Curve Analysis. We also discuss ideas for simplifying the inference procedure of an SCA and for potential additional inference on SCA results. We begin with a formalization of the SCA.

3.1 Formalized Procedure for Conducting an SCA

In this section, we attempt to formalize the proposed procedure of conducting an SCA based on the work by Simonsohn et al (Simonsohn et al., 2019). We consider a complete Specification Curve Analysis as consisting of 1) the construction and analysis of a single specification curve, and 2) an inferential test on the single specification curve. We start by formalizing the first part of a complete SCA.

Chapter 1 and 2 discussed the procedure to generate a specification curve. Here, we attempt to summarize the procedure in a more precise manner. Say the researchers have determined a specific research question of interest, the following steps are required to generate a specification curve:

- **Step 1: Determine Specifications** The first step is to determine the set of *specifications* based on reasonable alternative *operational decisions*. Such operational decisions may include:
  - Outlier decisions
  - Variable selection decisions
  - Variable transformation decisions
  - Interaction term decisions
  - ...

These operational decisions must be non-redundant, consistent with expertise in the field, and must not change the underlying theory being studied.

- **Step 2: Gather Estimates** Once the specifications are determined, the researcher runs all the reasonable models based on the set of specifications and gathers the point estimates from each model.
- **Step 3: Construct a Specification Curve** Once the point estimates are gathered, the researcher plots out a specification curve. The plot should include two parts: the upper plot including the curve of estimates, and the bottom plot indicating the specifications which produced the point estimates. The curve should also include information on the statistical significance of the point estimate, indicated it by color or size.
- **Step 4: Analysis of the Specification Curve** The analysis mainly encompasses two parts: determining evidence for a robust relationship in response to changes in specifications, and determining the existence of outstanding sets of specifications. In the case when a majority of the point estimates are of dominant sign and statistically significant, there exists evidence for a robust relationship in response to changes in specifications. If the majority of the point estimates are not statistically significant, or there appear to be similar numbers of positive and negative estimates, we conclude that there exists no evidence for a robust relationship in response to changes in specifications. If sets of specifications are producing an outstanding set of results, a closer look should be taken for patterns and the potential existence of unnoticed underlying theories. For example, if the set of specifications including operational decision D are mostly producing statistically significant estimates with the non-dominant sign, the set of specifications including D is considered an outstanding set of specifications and the researcher should examine if the inclusion of D implies a different underlying theory.

We now attempt to summarize and formalize the inferential tests that can be conducted on a specification curve. One of the inferential tests follows the structure of a hypothesis test and includes three main parts: identifying statistical hypothesis, determining test statistics and their reference null distributions, and inference on the test statistics. We formalize each of the three parts in the following sections.

3.1.1 Statistical Hypothesis

We first identify the statistical hypothesis. Consider the study is conducted to estimate some parameter θ. In general, the statistical hypotheses are of the following format:

- Null Hypothesis: $\theta = \theta_0$
- Alternative Hypothesis: $\theta \neq \theta_0$

Say that we want study the existence of a relationship between variables X and Y. We want to do so by estimating β where $Y_i = \beta_0 + \beta X_i + \gamma_1 Z_{1i} + ... + \gamma_k Z_{ki} + e_i$. Then the statistical hypotheses can be rephrased as the following:

- Null Hypothesis: No relationship exist, $\beta = 0$.
- Alternative Hypothesis: There exists a relationship, $\beta \neq 0$.

3.1.2 Test Statistics and Null Distribution

Generally a hypothesis test uses one test statistic. An SCA uses a set of three summary statistics of the specification curve as test statistics, and test the statistical significance using them jointly.

- **Summary Statistics** The set of test statistics consists of the following three summary statistics:
  - The median overall point estimate from the specification curve;
  - The share of estimates in specification curve that are of the dominant sign;
  - The share that is of the dominant sign and also statistically significant ($p < 0.05$).

To generate the null distribution, we use the resampling technique to produce a set of resampled data and construct specification curves on the resampled data. We have several options for the resampling technique. The most commonly known resampling technique that produces a null distribution is the *permutation resampling technique*. Say we want to study the relationship between Y and X, the technique would randomly reallocate observations of Y to observations of X. The randomness of allocation ensures that there exists no relationship between Y and X in the permutated dataset. Then the point estimate obtained from this permutated dataset is an estimate under the null hypothesis. The permutation technique would repeat the process a large number of times until we get an empirical distribution of estimates under the null hypotheses. In an SCA, with every resampled dataset, we compute estimates for all specifications and form them as a specification curve. We can thus obtain a null distribution of specification curves.

There exist other resampling techniques that also produce empirical null distributions. Simonsohn et al. mentioned using a bootstrapping technique in the case when there are no random assignments of treatment on subjects. We introduce one such technique when estimating regression coefficients (Bickel & Ren, 2001). To generate an empirical null distribution of $\hat{\beta}$, one first fit a regression model to obtain $\hat{(Y)}$, and obtain residuals $\hat{e} = Y = \hat{Y}$. One then sample the residuals with replacement from $\{\hat{e}_1, ..., \hat{e}_n\}$ to obtain $\{\hat{e}_1^*, ..., \hat{e}_n^*\}$, and define $Y_i^* = \hat{Y}_i + e_i^*$. Y^* will then be used as the response variable to estimate $\hat{\beta}*$. Repeat the resampling of residuals and computing $\hat{\beta}*$ many numbers of times produce an empirical distribution of $\hat{\beta}$. Similar as mentioned above, we can compute a specification curve for each resampled dataset to obtain the null distribution of specification curves.

3.1.3 Conclusion and Interpretation

The last step is to do inferences on the test statistics. Using the generated null distribution of the specification curves, we can obtain the null distribution of each of the summary statistics. For each specification curve in the empirical distribution generated as described in the previous section, we compute the three summary statistics. Gather the summary statistics from all specification curves in the empirical distribution, we obtain the null distributions of the summary statistics. We can then determine the p-values of our summary statistics by determining the percentile of our observed summary statistics in the null distributions. If the p-value is smaller than α, the summary statistic is statistically significant. The three summary statistics results are considered jointly for a conclusion.

This formalization of the SCA analysis and inference is based fully on the work and examples by Simonsohn et al. In the following section, we consider the possibility of modifying the process analytically and of additional inferences.

3.2 Implementing Theoretical Reference Distributions for the Summary Statistics?

In the case when there is a large set of specifications determined, generating an empirical null distribution of the specification curve may be computationally expensive. Would it be possible to determine theoretical reference distributions for the summary statistics? We propose ideas for two of the three statistics in the case when the point estimates are regression coefficient estimates $\hat{\beta}$.

3.2.1 Proportion of Point Estimates of Dominant Sign

For a specification curve, a set of β estimates is produced based on a set of different specifications. Let S be the set of specifications $S = S_1, S_2, ..., S_n$, where n refers to the total number of specifications. With each specification S_i, a model is run and an estimated regression coefficient $\hat{\beta_{S_i}}$ is obtained. The specification curve is formed by $\{\hat{\beta}_{S_1}, \hat{\beta}_{S_2}, ..., \hat{\beta}_{S_n}\}$. And the median overall point estimate in this case is $\tilde{\beta}_S = \text{median}[\hat{\beta}_{S_1}, \hat{\beta}_{S_2}, ..., \hat{\beta}_{S_n}]$. Let $T_{S_1}, T_{S_2}, ..., T_{S_n}$ be the indicator variables of the sign of the estimates. Say that $T_{S_i} = 1$ if $\hat{\beta_{S_i}}$ has dominant sign and equal 0 otherwise. Then under the null hypothesis, $\hat{\beta}$ follows a symmetric distribution centered at 0, so,

$$P[T_{S_i} = 1] = P[\hat{\beta_{S_i}} \text{ has dominant sign}] = 0.5$$

If $T_{S_1}, T_{S_2}, ..., T_{S_n}$ are iid, the sum of them follows a Bernoulli distribution with p =

0.5. The proportion of point estimates of dominant sign would then be $\frac{\sum_{i=1}^{n} T_{S_i}}{n}$, and we can compute:

$$P\left[\frac{\sum_{i=1}^{n} T_{S_i}}{n} = a\right] = P\left[\sum_{i=1}^{n} T_{S_i} = an\right] = \binom{n}{an} 0.5^n$$

This gives us the probability distribution of the proportion of point estimates of the dominant sign when there is no true relationship. It would then be possible to generate a null reference distribution analytically.

However, $T_{S_1}, T_{S_2}, ..., T_{S_n}$ are not iid distributed. The $\hat{\beta}_{S_i}$ estimates are correlated to each other, as they are estimated from similar models using the same dataset. The sign of these estimates are thus also correlated. In certain cases, the correlation can be determined. For example, Yitzhaki (Yitzhaki, 1990) provided a method to assess the sensitivity of a regression coefficient to monotonic transformations. It is shown that in some cases, no monotonic transformation can change the sign of the regression coefficient. One can also use the method proposed by Yitzhaki to find the type of monotonic transformation that does not change the sign of the regression coefficient. Say in an SCA, a monotonic transformation on one of the variables is determined to be a specification. There is a non-zero probability that the monotonic transformation would not change the sign of the regression coefficient. Let's call the two models, one using the transformation and one does not, S_i and S_j. If it is the case that the monotonic transformation does not change the sign of the regression coefficient, we would have $P[T_{S_i} = 0|T_{S_j} = 0] = 1$ and $P[T_{S_i} = 0|T_{S_j} = 1] = 0$. Under the null, it is true that $P[T_{S_i} = 0] = 0.5$. Clearly, $P[T_{S_i} = 0|T_{S_j} = 0] \neq P[T_{S_i} = 0]$ and $P[T_{S_i} = 0|T_{S_j} = 1] \neq P[T_{S_i} = 0]$. The two variables are thus not independent.

Depending on the choice of specifications, it is likely that $T_{S_1}, T_{S_2}, ..., T_{S_n}$ are not all independent of each other. In this case, the analytical approach we discussed would not be applicable. A potential method that could be applied in this case is a Poisson approximation of the sum of dependent Bernoulli variables (Chen, 1975). To obtain explicit bounds between the Poisson approximation and the distribution of the sum, a specific type of dependence between variables is required. It is required that the dependence between the random variables decreases as the distance between them increases. This would require us to understand how $T_{S_1}, T_{S_2}, ..., T_{S_n}$ are dependent on each other. Further studies will be needed to determining the correlation and dependence between variables $T_{S_1}, T_{S_2}, ..., T_{S_n}$. So far, the resampling technique is still the most practical approach.

3.2.2 Proportion of Point Estimates with Dominant Sign

The approach to take for the proportion of point estimates with dominant sign will be similar to the proportion of point estimates, except now the indicator variables will only equal to 1 if the estimate is of dominant sign and statistically significant.

Under the null, the probability of an estimate being of dominant sign and statistically significant will be $\frac{\alpha}{2}$. If the indicator variables are all iid, we can use a similar approach and conduct a probability function of the summary statistic. However, for the same reason as in the above case, the indicator variables will not be independent in most cases, and an analytical approach may be complicated.

As a result, it is difficult to compute the distributions of the two summary statistics analytically. Due to the correlation and dependence between regression coefficients, the resampling technique may indeed be the most practical approach as it mirrors the correlation found in the dataset. In the next section, we consider the validity of some additional inferences on the SCA.

3.3 Interpreting Numerical Values of Test Statistics

When studying the existence of an effect or a relationship, researchers care about the sign and magnitude of the effect or a relationship, if it exists. Normally, when studying such problems, only one estimate of the effect/relationship would be provided, and the magnitude and sign of the estimate are considered meaningful. In the case of SCA, however, any of the point estimates on the curve would have been a reasonable estimate of the effect/relationship, since they are generated from a reasonable model that could have been analyzed individually. These estimates may have different signs and may have different magnitude. While the inference on a specification curve answers the question of whether or not there is evidence for the existence of an effect/a relationship, it does not give an exact estimate of the effect/relationship. Can we combine the point estimates in a specification curve and generate an overall estimate of the effect/relationship?

As mentioned earlier, Orben used the median of the point estimates in a specification curve to represent the overall estimate of the effect/relationship. When we have a set of estimates of the same thing, it is tempting to use the center of these estimates as representing the overall result. The median of them seems not to be a bad choice. However, when the numbers in a sample are of different scales, any summary statistic is meaningless. Certain choices of specifications may lead to differences in model forms, and the numerical values of different point estimates may have different scales. For example, Simonsohn et al. in one of their examples considered using the log transformation of the response variable as an alternative specification to using the response variable itself. For those point estimates generated with “log transformation”, we should interpret the numerical values as the changes in the logged response variable when the independent variable changes by one unit. For those point estimates generated using the response variable itself, the interpretation of the numerical values would be the changes in the response variable when the independent variable changes by one unit. A large point estimate in the latter case does not necessarily reflect a stronger

effect/relationship than a small point estimate in the former case, as the numbers are of different scales.

It is possible that variable transformation is included in the set of specifications. Does this mean that the point estimates can be interpreted in the same way? Not necessarily. If the inclusion of some interaction term is determined to be a specification, the way of interpreting the β estimate, in this case, will be different from the interpretation of it without the interaction term. Moreover, decisions that change the control variables being considered in the model would also produce point estimates being interpreted in different ways. For example, when the control variables are A, B, and C, the way one would interpret β estimate is that it represents the effect/relationship when A, B, and C are controlled. When the control variables are different, say A and B, the point estimate represents the effect/relationship when A and B are controlled. The estimates have different meanings.

As different point estimates may be interpreted differently, the median point estimate may not represent the median estimated effect/relationship. With the great flexibility of choosing specifications in SCA, the different point estimates should likely be interpreted differently. For example, in the case of Orben's work, even if the type of specifications used is limited to three types, the different point estimates correspond to the different independent variables, different response variable, and different set of control variables, with many of the variables having different scales. Especially in the case when the alternative independent variables have different scales, it is inappropriate to interpret the different point estimates in the same way and consider the median point estimate as representing the median estimated effect/relationship. We would have to restrict the types of specifications to a very small set if we want the summary statistic of the point estimates to be meaningful. But then we lose the flexibility in operational decisions of an SCA.

Instead of computing the summary statistics directly from the point estimates, Bayesian Model Averaging (See (Hoeting, Madigan, Raftery, & Volinsky, 1999) for an overview) may allow us to obtain numerically meaningful estimates. The Bayesian Model Averaging method considers a prior set of reasonable models, similar to the idea of the set of specifications in an SCA. BMA would compute posterior model probabilities on each model, which is the probability of each model being a good fit given the data. Averaging over the posterior distributions under each of the models will then produce the posterior distribution of the overall point estimate given data. The summary statistics of this posterior distribution would provide reliable numerical estimates. BMA can be implemented in several different cases, including linear regressions with differences in predictors, outliers, and transformations; generalized linear regression with changes in the choice of the independent variables, the link function, and the variance function; etc.

One direct approach to obtain an overall point estimate of interest would be applying BMA on the models constructed based on the specifications. In Orben's study, the model form is restricted to linear regression, and the specifications are restricted to differents in predictors and the response variables. In this case, the existing Bayesian

Model Averaging method on linear regressions (Hoeting et al., 1999) can be directly implemented. The idea of BMA also provides a direction for future improvement of the SCA method, in which an overall estimate may be possible to obtain.

We have now formalized the SCA procedure and discussed potential direction and ideas of developing the method. In the Appendix, we implement the formalized SCA procedure to illustrate a "corrected" application on the research topic "*the relationship between digital technology use and teen mental well-being*".

Conclusion

The Specification Curve Analysis allows researchers to assess the robustness of model estimates in response to changes in operational decisions. An inferential test on the specification curve tests the existence of a non-zero relationship/effect and provides information on the direction of it. While there is great flexibility in the choices of specifications, the SCA can only work with a valid set of reasonable operational decisions for the specific research question. In the case when the method is applied inappropriately, the reliability of the study results can be challenged.

Orben's application of SCA consists of several problems that challenge the reliability of its results. The study originally concluded that a negative effect of technology use on teenagers' mental well-being exists, but the magnitude of the effect is too small to warrant policy changes. It is possible that after fixing the problems in this application, the original results and conclusion will be overthrown, but we can get one step closer in understanding the true effect of digital technology use on teenagers' mental well-being.

We propose several ideas on improving the SCA, including the attempt to determine analytical reference distribution for test statistics, and the attempt to obtain estimates which has interpretable numerical values. The great flexibility of the types of specifications brings excessive difficulty in analyzing the reference distribution analytically. While the summary statistics of the estimates from a specification curve is not interpretable, an overall point estimate may be computable using methods such as Bayesian Model Averaging. Due to time constraint, the ideas have not been rigorously proven or tested, but can be good start points for future studies on SCA.

Appendix A

Re-analysis of YRBS Data on Technology Use vs. Teenager Mental Well-Being

We have identified the problems existing in Orben's application of SCA when studying the relationship between digital technology use and teenager mental well-being. In this appendix, we attempt to illustrate an SCA application following the formalized procedure on the same topic, using one of the three datasets. However, this application is conducted without expert knowledge in the field of Psychology. The specifications determined in this study may not be valid from the perspective of an expert in Psychology. The application should only be considered as an illustration and should not be considered as a serious Psychology study.

A.1 Research Question of Interest and Specifications

We first choose our specific research question of interest, the relationship between TV use and teenager mental well-being. We choose the dataset YRBS (Kann L, 2015) and regression models to study the question. Based on the research question of interest and variables available in the data, we specify the independent variable to be the survey answer to the question:

- "On an average school day, how many hours do you watch TV?"

Orben used four variables to compute alternative measures to the dependent variable, teenager mental well-being. Without better expertise in the field, we follow Orben and consider the four choices of the response variable as part of the operational decisions we determine. The four variables are the answers to the questions:

- "During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?"
- "During the past 12 months, did you ever seriously consider attempting suicide?"
- "During the past 12 months, did you make a plan about how you would attempt suicide?"
- "During the past 12 months, how many times did you actually attempt suicide?"

The fourth variable, although was not designed as a binary variable, has only two values: 0 and 1. Thus, we consider an alternative interpretation of the fourth variable as, "during the past 12 months, did you actually attempt suicide?", which makes it a binary variable in this case. Since three of the response variables are binary variables and the fourth variable can be reinterpreted as a binary variable, a logit regression model may be appropriate. Orben chose to use linear regression on the same set of response variables. In this application, we consider both model forms as appropriate potential model form and consider the specification of:

- Logit regression model
- linear regression model

I would consider two control variables as necessary for the models, "On an average school day, how many hours do you play video or computer games or use a computer for something that is not school work?" and "During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?". The choices are not supported by Psychological theories but are used here to illustrate the process.

I then determined two other control variables which can be included in the models but do not have to be:

- "During the past 30 days, on how many days did you have at least one drink of alcohol?"
- "During your life, how many times have you used hallucinogenic drugs, such as LSD, acid, PCP, angel dust, mescaline, or mushrooms?"

Specifications considered here will include specification without the two control variables, with one of the two variables, and with both variables.

The last specification considered is the inclusion/exclusion of an interaction term between one of the control variables and the independent variable.

Combining all the specifications mentioned, 128 models have been specified and ran for analysis.

A.2 Specification Curve Results and Analysis

Running the 128 models produce the following specification curve plot:

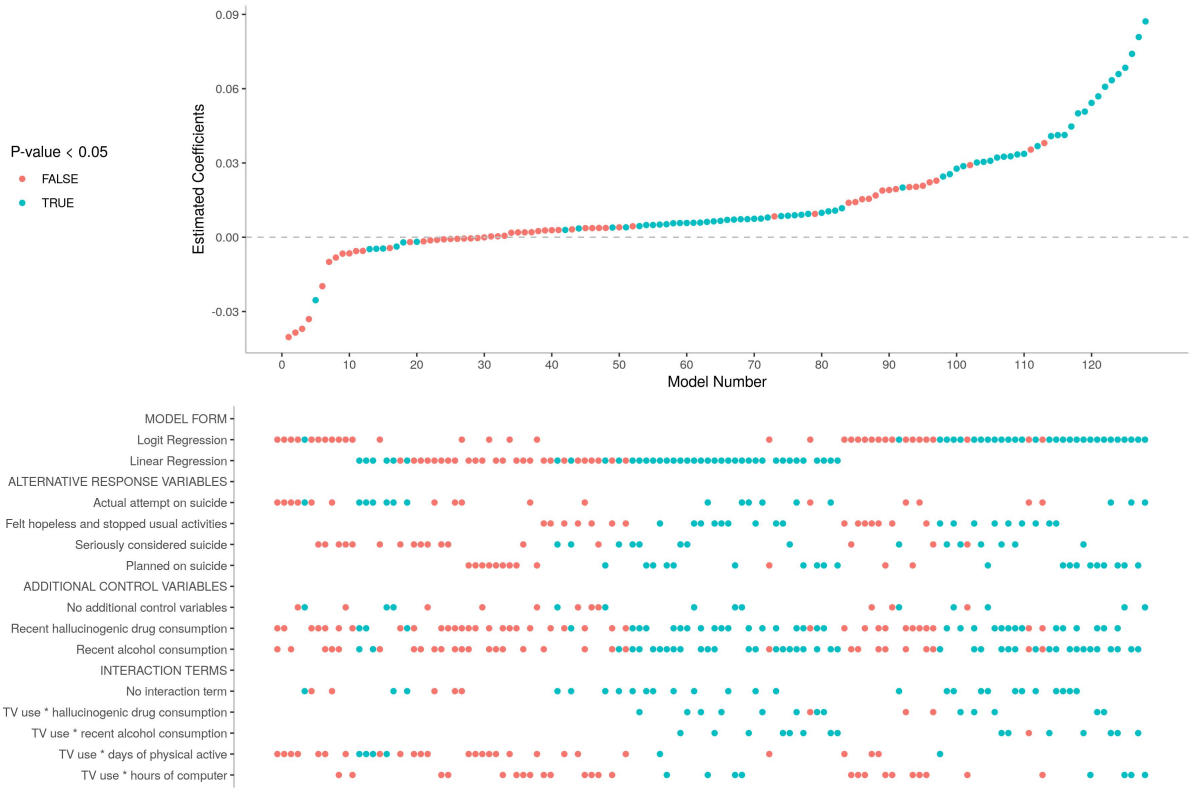

Figure A.1: The specification curve of 128 specifications. Blue dots represent statistically significant estimates, while the red dots represent insignificant estimates

As we can observe from the curve, the majority (97 out of 128) of the estimates are positive. 69 of the estimates are statistically significant, with 62 of the estimates being positive and statistically significant. The majority of the statistically significant results are positive. There appears to be evidence for a positive significant effect, and the effect appears to be moderately robust in response to the changes in specifications.

We then look for the existence of an outstanding set of specifications in the plot. Of the models that included the interaction term between TV use and alcohol consumption, the estimated coefficients are all positive, and only one of the estimated coefficient is not statistically significant. For those specifications included the alcohol consumption variable but not the interaction term, the majority of the estimates are not statistically significant. This may imply a dependence between TV use and alcohol consumption, where the effect of TV use on teenagers' well-being can be different depending on alcohol consumption.

A.3 Inferential Test Results and Analysis

The last part of an SCA is an inferential test. We conducted a permutation test with 500 resampled datasets. We compute only the inference on the summary statistics method.

We obtained the median point estimate of the SCA to be 0.006514, the proportion of positive estimates to be 0.7656, and the proportion of positive significant estimates to be 0.4844. Based on the permutation distribution, we found the p-values to be: 0.002 for the median, 0.146 for the proportion of positive estimates, and 0 for the proportion of positive significant estimates. Two of the test statistics are found to be statistically significant. This suggests evidence for the existence of a positive effect of TV use on teenagers' mental well-being, and we reject the null hypothesis.

References

Bessiere, K., Kiesler, S., Kraut, R., & Boneva, B. S. (2008). Effects of internet use and social resources on changes in depression. *Information, Community & Society*, *11*(1), 47–70.

Bickel, P. J., & Ren, J.-J. (2001). The bootstrap in hypothesis testing. *Lecture Notes-Monograph Series*, 91–112.

Burke, M., Kraut, R., & Marlow, C. (2011). Social capital on facebook: Differentiating uses and users. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 571–580).

Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *The Annals of Probability*, *3*(3), 534–545. Retrieved from `http://www.jstor.org/stable/2959474`

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 382–401.

Johnston, L. D., Bachman, J. G., O'Malley, P. M., Schulenberg, J. E., & Miech, R. A. (2017). Monitoring the future: A continuing study of american youth (8th- and 10th-grade surveys), 2016. Inter-university Consortium for Political; Social Research [distributor]. `http://doi.org/10.3886/ICPSR36799.v1`

Kann L, H. W., McManus T. (2015). Youth risk behavior surveillance — united states, 2015. MMWR Surveill Summ 2016;65(No. SS-6):1–174. http://doi.org/http://dx.doi.org/10.15585/mmwr.ss6506a1

Orben, A. (2019a). NHB_2019. *GitHub repository*. `https://github.com/OrbenAmy/NHB_2019`; GitHub.

Orben, A. (2019b, September). A journal club to fix science. *Nature News*. Nature Publishing Group. Retrieved from `https://www.nature.com/articles/d41586-019-02842-8`

Orben, A. K., A. & Baukney-Przybylski. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, 173–182.

Orben, A., & Przybylski, A. K. (2019). Screens, teens, and psychological well-being:

Evidence from three time-use-diary studies. *Psychological Science*, *30*(5), 682–696. http://doi.org/10.1177/0956797619830329

Orben, A., & Przybylski, A. K. (2020, January). Analysis code. OSF. Retrieved from osf.io/e84xu

Orben, A., Dienlin, T., & Przybylski, A. K. (2019). Social medias enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences*, *116*(21), 10226–10228. http://doi.org/10.1073/pnas.1902058116

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science*, *28*(12), 1821–1832. http://doi.org/10.1177/0956797617723726

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2019). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *Available at SSRN 2694998*.

University of London, C. for L. S., Institute of Education. (n.d.). Millennium cohort study: Sixth survey, 2015 [data collection]. 5th Edition. UK Data Service. http://doi.org/http://doi.org/10.5255/UKDA-SN-8156-5

Verduyn, P., Ybarra, O., Résibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review*, *11*(1), 274–302.

Yitzhaki, S. (1990). On the sensitivity of a regression coefficient to monotonic transformations. *Econometric Theory*, *6*(2), 165–169.