My Final College Paper

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 want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

SCA and Its Applications

In this chapter, we focus on the Specification-Curve Analysis proposed by Simonsohn et al. [cite] and its existing applications. We discuss the details of conducting an appropriate SCA and then provide a description of two existing applications of SCA.

1.1 Specification-Curve Analysis

For the following sections, we will follow closely the form of specification-curve analysis as proposed by Simonsohn et al. [cite] By Simonsohn, conducting a specification-curve analysis involves three steps: (1) Identifying the set of specifications, (2) Estimate overll all 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 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 *operationalization 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 intersection 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 variable “hair color appearance” will be an inappropriate specification to use for conducting an SCA on this research question.

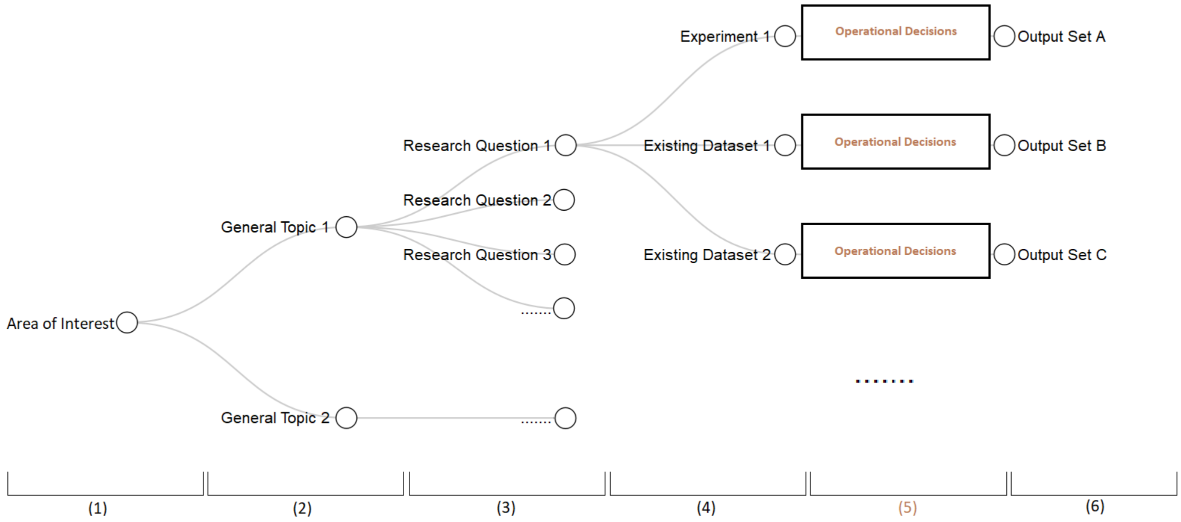


Figure 1.1: Visualization of scientific study steps

A visualization may help with understanding this idea of operational specifications. 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.

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 specifications inside one of the boxes of operational decisions are varied, 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 are still considered reasonable. 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 which could be generated by choosing different combinations of the operational decisions. A visualization of this process is shown below:

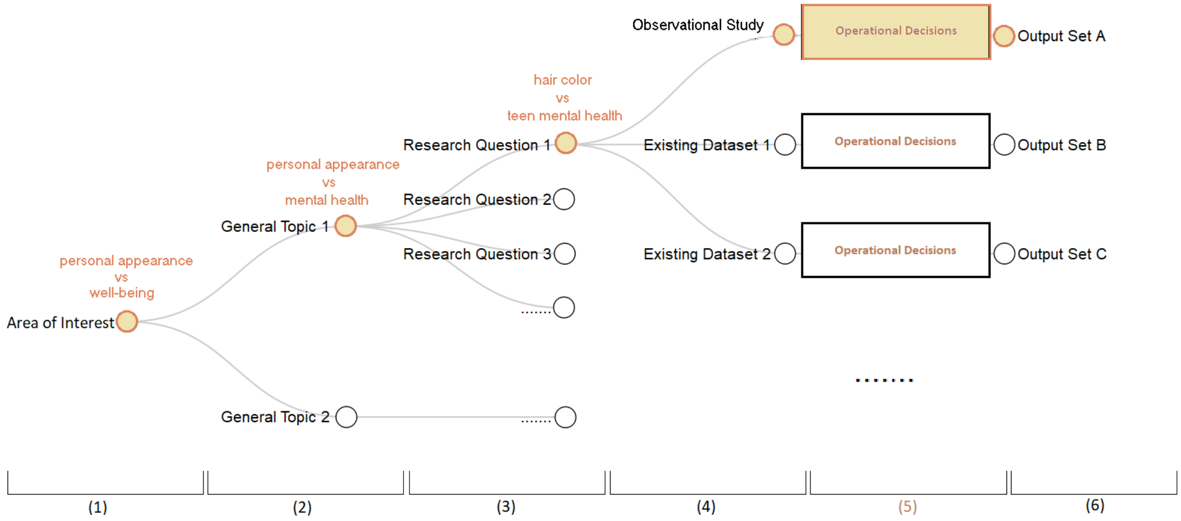


Figure 1.2: Procedure of a Scientific Study

Commonly in practice, the psychologist chooses one certain specification and produce the estimate from there. An SCA would consider all reasonable specifications and the full set of reasonable outputs. The following figure provides a closer visualization on the operational decisions:

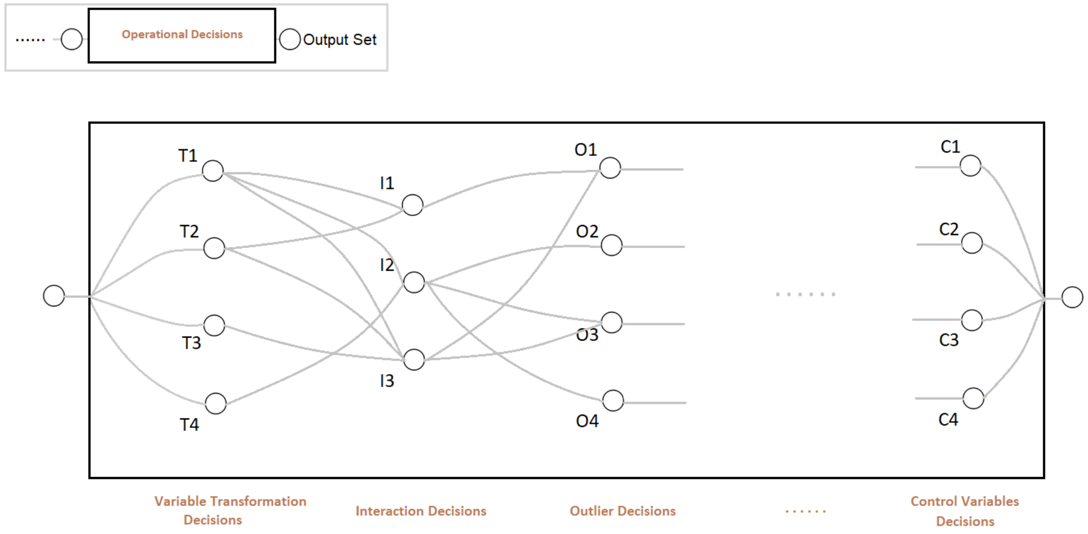


Figure 1.3: Visualization of Operational Decisions

In this figure, each node represents a unique decision of the type that could be made by the researchers, and the branches connect appropriate combinations of decisions. Following the branches in different ways can lead to different combinations of these operational decisions, and will produce sets of unique models which then produce a set of possible model outputs. In common practice, a researcher may choose one specifc 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 not be considered outliers anymore 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 specifications. But among the existing applications of SCA (see Section 1.1.2 for details), after enumerating the specifications of each type, all combinations of the enumerated specifications are considered, which can be visualized as in the following figure:

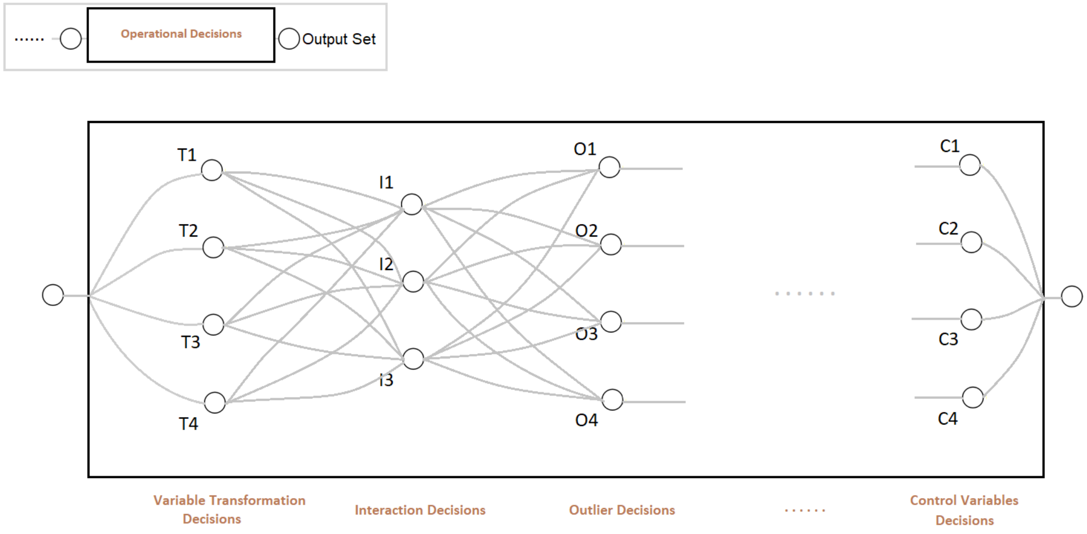


Figure 1.4: Visualization of Operational Decisions considered in existing applications

1.1.2 Constructing Specification Curve

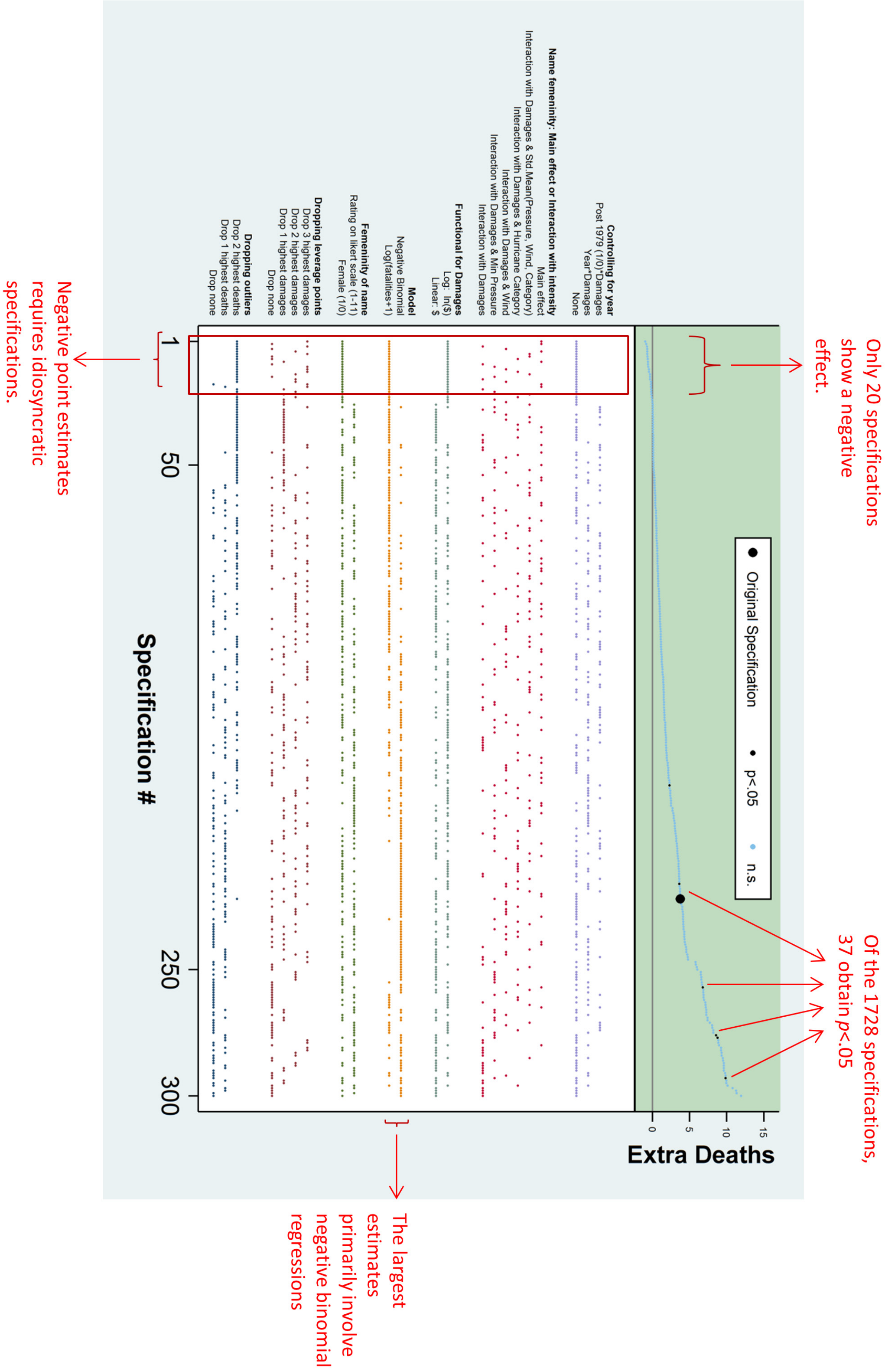
The next step is to build a specification curve. After determining the specifications, a set of the specifications can be determined, where each specification leads to a different model to be run. Here we describe how the existing applications tend to generate the set of specifications. Say a group of 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 specifications and will produce models 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 awhile 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 brings computational obstacles, a random subset of the specification models can be used instead.

Now that all the models have been determined, the next step is to estimateall of the models and extract the point estimates from each of the models. In the case of linear regressions, the extracted point estimates are generally the β estimates from 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 set of specifications that generated the specific model for this estimate.



Only 20 specifications show a negative effect.

Of the 1728 specifications, 37 obtain $p<.05$

Negative point estimates requires idiosyncratic specifications.

The largest estimates primarily involve negative binomial regressions

Figure 1.5: Specification Curve

As shown in Figure 1.5, 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 lowest value to highest value. Each model is produced by a specific set of specfications. The vertical axis is the numerical value for the estimates, and the values on the horizontal axis represent the set of decisions determined for that model, represented by dots in the bottom plot. In the bottom half of the plot, each dot represents the various decisions of that model. 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 plot 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 purely 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.

1.1.3 Analysis on Specification Curve

The last step of an SCA is the statistical inference on the single specification curve result. The question for an inferential analysis, as stated by Simonsohn [cite], is "*Considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?*" The authors suggested that while the question is hard to be answered analytically, one can generate an expected distribution of the specification curves when the null hypothesis is true 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 assignment.

Once a 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 that can be used for the inferential analysis is a "confidence interval" of the specification curve. Consider in the case of estimating regression coefficients. For each β estimate, using the null distribution generated by resampling, we can compute a confidence interval of the significant level desired. The confidence interval of the specification curve would be two curves, one formed by the lower bound of the confidence intervals for individual estimates, and the other one formed by the upper bound of the confidence intervals for individual estimates. It's claimed in the example provided by Simonsohn et al. that, in the case when the specification curve falls completely outside of the "confidence interval", i.e. there is no overlap between the curves, the null is rejected.

Using the test statistics in combination with the confidence interval curves would provide the answer to our question of interest. The followings are two examples given by Simonsohn et. al:

Table 1.1: (#tab:ex_test_stat) The test statistics results from two of the examples as provided by Simonsohn, where 2b) studied whether or not resumes with distinctively Black names ower callback rates when applying jobs, and 2c) studied whether or not resumes with distinctively Blac 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

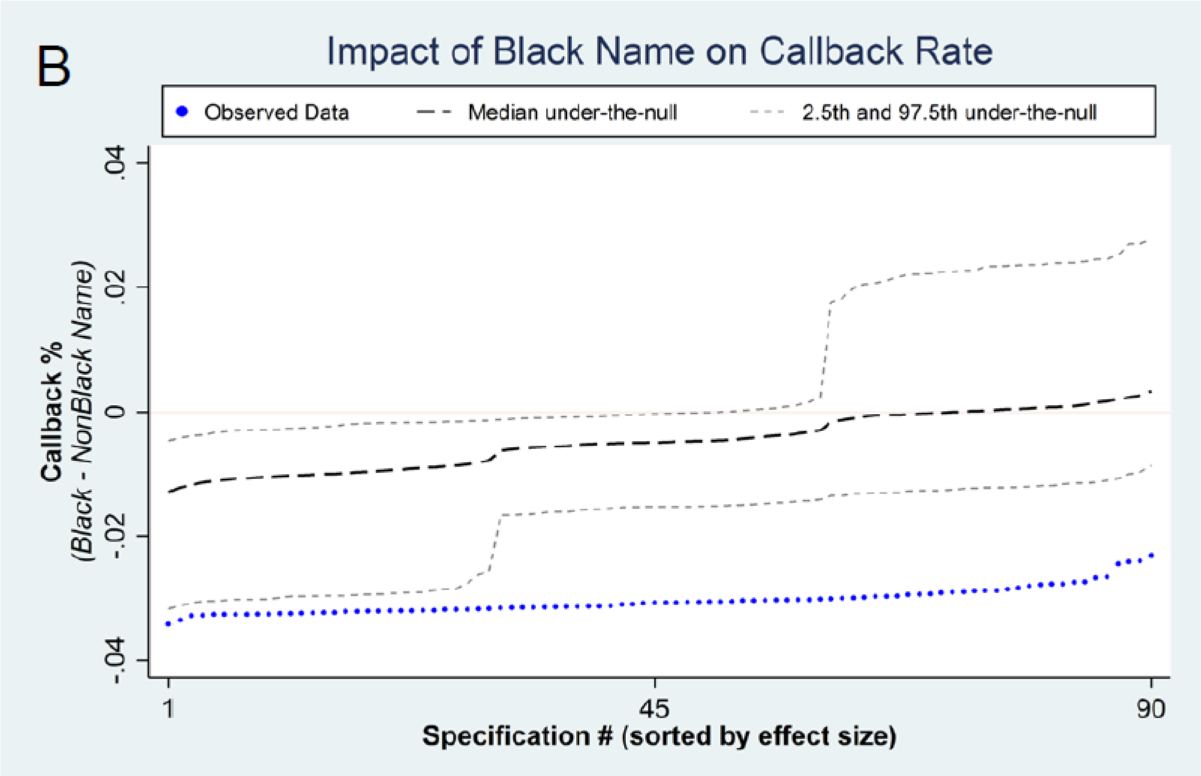


Figure 1.6: Confidence interval of the specification curves

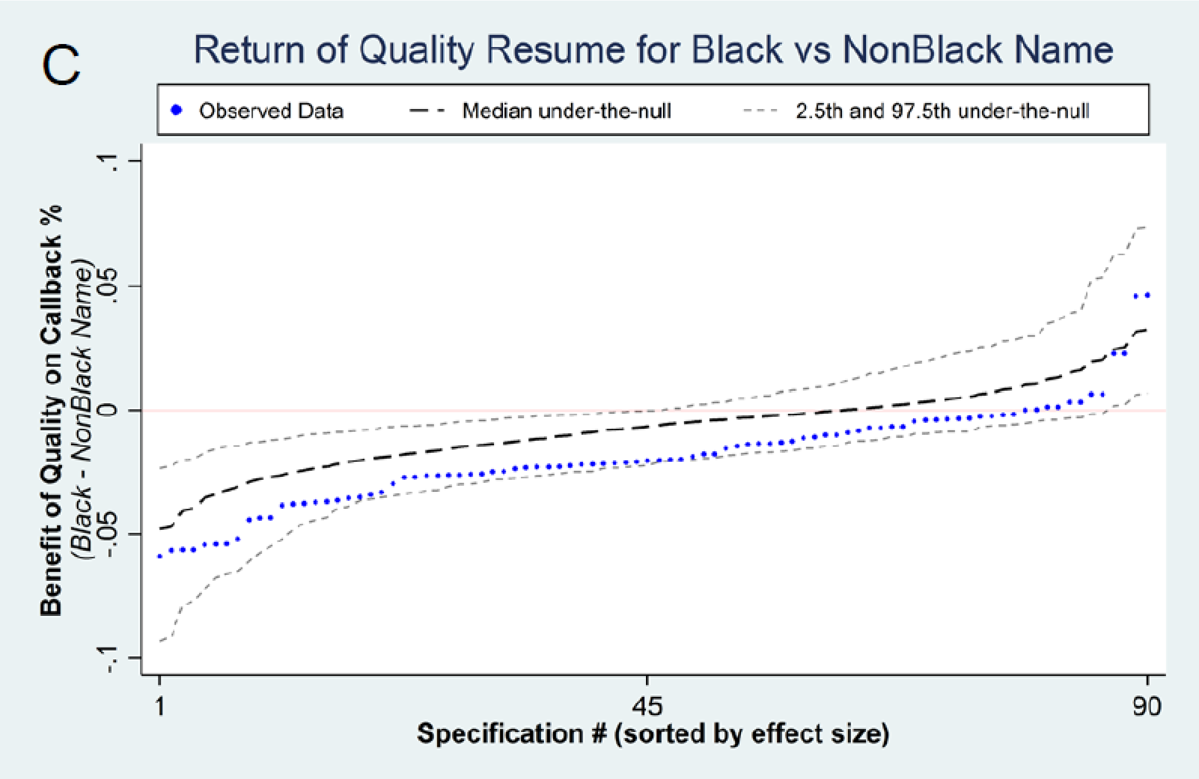


Figure 1.7: Confidence interval of the specification curves

> 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. In the other case, the curve clearly has 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. The authors did not specify whether or not the two methods, using summary statistics as test statistics or using the confidence interval curves, must be used in combination for a complete SCA inference test or can be used separately. The existing applications of the SCA, however, tend to use only the test statistics methods and not the confidence interval curves. In the following sections, we introduce the existing applications of SCA method with emphasis on two Psychology applications. A more detailed discussion of the applications will be provided 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), 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), Youth Risk and Behaviour Survey (YRBS), and Millennium Cohort Study (MCS). 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. 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. 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 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 alternate to each other. One example will be: 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

This section discusses the attempt to replicate Orben's study along with the assessment of its use of SCA. We start from a more detailed introduction of the study and its aims, and then we discuss the details in the study. We will introduce the three datasets used and the attempt to replicate the study following Orben's procedure, including the obstacles to overcome during the replication process.

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 being made for reproducibility. Without these efforts, this replication would not be possible.

*

The association between adolescent well-being and digital technology use* was published in 2019 on the journal Nature Human Behaviour [cite]. It is an online journal which publishes "research of outstanding significance into any aspect of individual or collective human behaviour". 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 researches published on the journal often are those having outstanding findings and produce great influece in related fields. Orben's study, being published on January 2019, has already had 152 citations (as by April 2020), and is bringing great influence to the field of Psychology.

Tremendous efforts have been made by Dr. Orben to achieve reproducibility of this study. 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 [cite]. 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 possible. This thesis would not have been done so 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 [Cite Nature article]

The solid aim for openness and reproducibility shown by Dr. Orben greatly encouraged the completion of this thesis. We value the great 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 advices 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) from the United Kingdom, Youth Risk and Behavior Survey (YRBS) from the US, and Monitoring the Future (MTF) 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 [cite] 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 [cite]. 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 [cite] 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 [cite]. 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". [page 8 in Orben's, cite] 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 avaible in format including SAS, SPSS, Stata etc, but were only avaiable as separated sub data files. After Rstudio, the data can be obtained from one complete R datafile for easy implement 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 [cite] 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 from [cite]. Access to the data is open to registered users of UK Data Service with 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

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 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 on OSF, 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 in more reasonable, 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 $Median(\beta) = -0.035$ 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 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, proportion of estimates with dominant sign, and 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. With access to an AWS server with 96 cores, the running time can be further reduced. 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 of no effect.

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 platform, 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 teenager mental well-being, or could the different types of digital technologies be actually having different relationships with teenager 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 internet may account for different outcomes in measure of depression. (Bessiere, Kiesler, Kraut, & Boneva, 2008) [cite] Burke et. al looked more specifically on how using internet passively or actively can result in different outcomes in social communication skills and self-esteem. (Burke, Kraut, & Marlow, 2011) [cite] Similar results were also obtained by Verduyn et al. (Verduyn, Ybarra, Resibois, Jonides & Kross, 2017) [cite]. The existing literature suggests considering the active usage of medias and passive usage of medias as having different effect on mental health and other personality scales. The different types of technologies allow 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 interchangable 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 the digital technology, based on the existing literature, the usage of different technologies may have 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 1.4. 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.

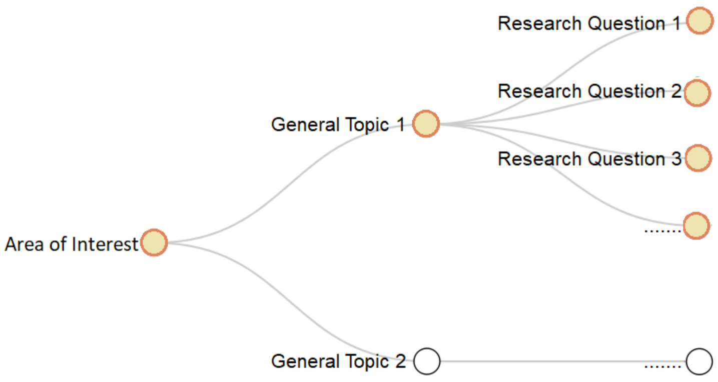


Figure 2.1: Orben considers multiple research questions as interchangeable with one another instead of choosing a specific research question from the general topic, contrary to the recommendation of Simonsohn et al. [cite]

Such 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 all specifications relating to the inclusion/exclusion of variables. The determined specifications can be 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.

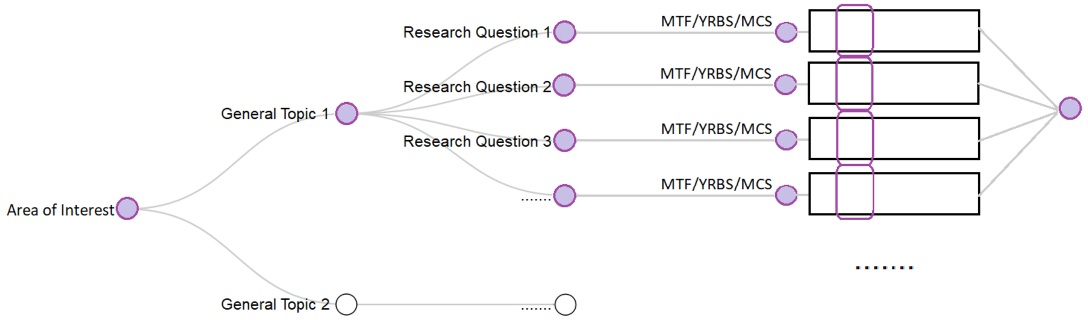


Figure 2.2: The set of specifications considered by Orben

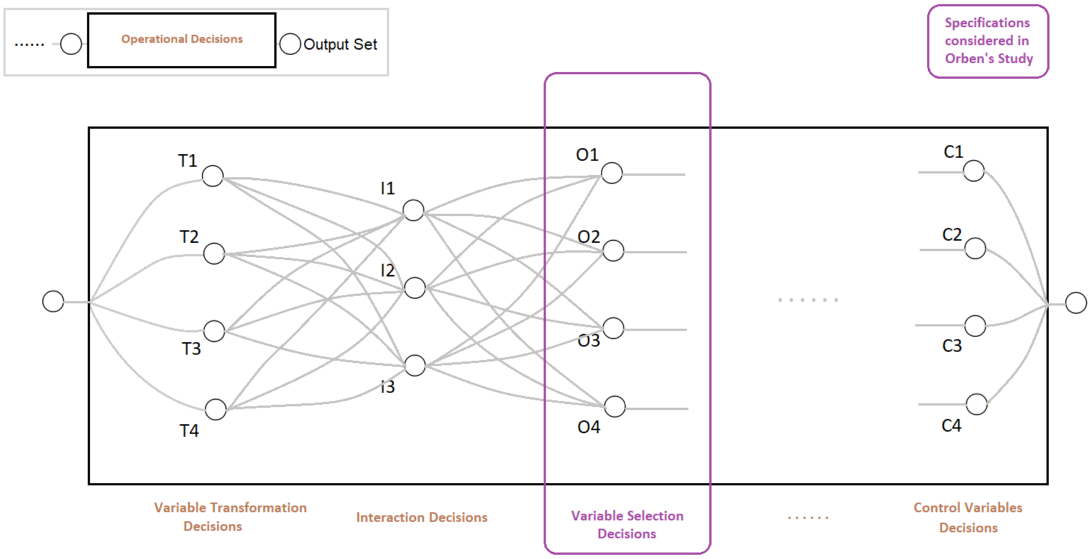

......
Operational Decisions
Output Set
Specifications considered in Orben's Study
T1
T2
T3
T4
I1
I2
I3
O1
O2
O3
O4
......
C1
C2
C3
C4
Variable Transformation Decisions
Interaction Decisions
Variable Selection Decisions
......
Control Variables Decisions

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 conducting an analysis for 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 exist 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 uncatched 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, following the instructions suggested by Simonsohn et al. 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. [cite] may in fact provide unreliable results and interpretations.

Chapter 3

Formalization and Further Ideas on 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 from a formalization of the Specification Curve Analysis.

3.1 Formalization of SCA

In this section, we attempt to formalize the proposed procedure of conducting an SCA based on the work by Simonsohn et al [cite]. We consider a complete Specification Curve Analysis as consisting: 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, constructing and analyzing a single specification curve.

3.1.1 Formalizing a Single Specification Curve

Chapter 1 and 2 has provided discussions on the procedure to generate a specification curve. Here, we attempt to summarize the procedure to construct a specification curve. 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

  - . . . The requirements of these operational decisions is that they are valid and non-redundant operational decisions based on the field expertise, and making the decisions do not change the underlying theory being studied.

- [**Step 2: Gather estimates**] Once the specifications are determined, the researcher run all the reasonable models based on the set of specifications and gather 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 curve of estimates, and the subplot indicating the specifications which produced the point estimates. The curve should also include information on the statistical significance of the point estimate, indicate it by color or size.

- [**Step 4: Analysis of the Specification Curve**] Now the researcher analyzes the specification curve. The analysis encompasses two parts: determining evidence for a robust relationship in response to changes in specifications, and determining 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 appears 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 ba taken onto the sets of specifications for patterns and potential implication of uncatched underlying theories. For example, if a set of specifications both include operational decision D, and they are all producing statistically significant estimates of the non-dominant sign, the set of specifications is an outstanding set of specifications and the researcher should examine if the inclusion of D may have catched a different underlying theory.

3.2 Formalization of Inferential Test on Specification Curve

We now attempt to summarize and formalize the inferential test on a specification curve. The inferential test is in the structure of a hypothesis testing, including three main parts: identifying statistical hypothesis, determining test statistics and reference null distribution, and inference on the test statistics. We formalize each of the three parts in the following sections.

3.2.1 Statistical Hypothesis

- Null Hypothesis: No effect/relationship exist.
- Alternative Hypothesis: There exists a non-zero effect/relationship.

3.2.2 Test Statistics and Null Distribution

Generally a hypothesis test uses one test statistics. An SCA uses two sets of "test statistics": specification curve its self and the summary statistics of the specification curve. We describe the two "test statistics" separately:

- [**Speficiation Curve**] The first "test statistic" is the specification curve itself, which encompasses point estimates determined from different specifications. The reference null distribution will be a distribution of the specification curves under the null.
- [**Summary Statistics**] The next set of "test statistics" is the three summary statistics as suggested:
  - 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. For each resampled data, we can produce a specification curve and compute the summary statistics.

3.2.3 Conclusion and Interpretation

The last step is to do inferences on the test statistics. We describe the inferences for the two sets of test statistics separately:

- [**Inference on the specification curve**] Using the generated null distribution of the specification curves, we compute the 95% "confidence interval curves". The confidence interval curves includes two curves: the lower bound curve and the upper bound curve. The lower bound curve is the curve formed by the 2.5% quantile of each of the point estimates, and the upper bound curve is the curve formed by the 97.5% quantile of each of the point estimates.
- [**Inference on the summary statistics**] Using the generated null distribution of the specification curves, we can compute the null distribution of each of the summary statistics by computing them on each specification curve in the null distribution. We then determine the p-values of our summary statistics with reference to their 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.
- If the specification curve falls completely outside of the confidence interval curves, and if the majority of the summary statistics are statistically significant, we reject the null hypothesis of no relationship/effect.

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

3.3 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 consider the three statistics separately in the case when the point estimates are regression coefficient estimates $\hat{\beta}$.

3.3.1 Median Overall Point Estimate

For a specification curve, a set of β estimates is produced based on different specifications. Let's call the specifications $S_1, S_2, ..., S_n$, where n refers to the total number of estimates. 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 $\tilde{\beta_S} = \text{median}[\hat{\beta_{S_1}}, \hat{\beta_{S_2}}, ..., \hat{\beta_{S_n}}]$. Is it possible to determine the distribution of $\tilde{\beta}$ under the null, when there is no relationship?

When there is no true relationship, each of the $\hat{\beta_{S_i}}$ follow a normal distribution centered at 0. The variance of the $\hat{\beta_{S_i}}$ may differ. The $\hat{\beta_{S_i}}$ will be correlated to each other, as they are estimated from similar models using the same dataset. If we consider each β_{S_i} as a normally distributed random variable that are correlated to each other, the sample of $\hat{\beta_{S_1}}, \hat{\beta_{S_2}}, ..., \hat{\beta_{S_n}}$ can be considered as an observation from a multivariate normal distribution. If we can determine the covariance matrix for this multivariate normal distribution, we may be able to get one step closer into computing the distribution of the median estimate analytically.

Is it possible to determine the covariance matrix of $\hat{\beta_{S_1}}, \hat{\beta_{S_2}}, ..., \hat{\beta_{S_n}}$? An answer to the question has not yet been found. Say if it is possible to compute the covariance between any two $\hat{\beta_{S_i}}$ and $\hat{\beta_{S_j}}$ $(i \neq j)$. In the case when there is a large number of specifications determined, the dimension of the covariance matrix will be huge and it could be difficult working with a matrix of high dimensions. For practical application purpose, using the resampling technique to generate a reference null distribution may indeed be more straightforward and easy to adapt.

3.3.2 Proportion of Point Estimates of Dominant Sign

Let's consider again the point estimates in a specification curve $\hat{\beta_{S_1}}, \hat{\beta_{S_2}}, ..., \hat{\beta_{S_n}}$. 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. Then under the null of no relationship, we have

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

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 determine:

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

However, are $T_{S_1}, T_{S_2}, ..., T_{S_n}$ iid? The answer might be no. For example, Yitzhaki [cite] provided a method to assess the sensitity 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 do 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. Then in this case, $P[T_{S_i} = T_{S_j}] > 0$. Under the null, it is true that $P[T_{S_i} = 0] = 0.5$. But $P[T_{S_i} = 0|T_{S_j} = 0] = 1$ and $P[T_{S_i} = 0|T_{S_j} = 1] = 0$. The two variables are thus not independent.

Depending on the choice of specifications, it is possible that $T_{S_1}, T_{S_2}, ..., T_{S_n}$ are not all independent of each other. When the set of specifications is large, one would have to prove a great number of independence between variables before able to apply the pobability function mentioned. Similar to the median point estimate, for practical application purpose, using a resampling technique may be more straightforward and easy to adapt.

3.3.3 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 statitistically 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 the similar approach and to conduct a probability function of the summary statistic. However, for same reason as in the above case, the indicator variables may not be independent in certain cases. For practical application purpose, using a resampling technique may be more straightforward and easy to adapt.

As a result, it is difficult to compute the distributions of the summary statistics analytically. Due to the correlation and dependence between variables, the resampling technique may indeed be more straightforward and easy to adapt. In the next section, we consider the validity of an additional inference on the SCA.

3.4 Interpreting Numerical Values of Test Statistics?

When studying the existence of an effect or a relationship, researchers almost always care about the sign and magnitude of the effect or a relationship, if 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, all the point estimate would have been a reasonble estimate of the effect/relationship. 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 estimate 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, the center or other summary statistic is unmeaningful. 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 alternative specification to using the response variable itself. For those point estimates generated with "log transformation" as one of the specifications, we should interpret the numerical values as the amount of changes in logged response variable when 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 amount of changes in response variable when independent variable changes by one unit. A large point estimate in the later case does not necessarily reflect a stronger effect/relationship than a small point estimate in the former case, as the numbers represent different type of effect/relationship between the two variables.

In many cases, decisions that change the model form are not considered reasonable alternative specification to each other. 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 interpreting the β estimate in this case will be different from the interpretation of it without the interaction term. Moreover, decisions 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. They do mean different things.

As different point estimate 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, it is likely that the different point estimate should 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 estimate correspond to different independent variable, different response variable, and different set of control variables, with many of the variables having different scale. It is inappropriate to interpret the different point estimate in the same way, and consider that the median point estimate represents 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 loose the flexibility of an SCA.

Instead of computing the summary statistics directly from the point estimates, Bayesian Model Averaging would allow us to obtain numerically meaningful estimates. The Bayesian Model Averaging method considers a prior set of reasonable models to run, similar as the idea of the set of specifications in an SCA. BMA would compute posterior model probability 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 model will then produce the posterior distribution of the overall point estimate of interest given data. The summary statistics of this posterior distribution would provide reliable numerical estimates. BMA has been implemented in several cases, including linear regressions with difference on predictors, outliers and transformations, generalized linear regression with changes on choice of the independent variables, the link function and the variance function, etc. For future work on further developing the SCA method, the implementation of Bayesian Model Averaging may be considered.

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

Chapter 4

Application of Formalized SCA

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 chapter, 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 expertise 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.

4.1 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 will be using the dataset YRBS [cite] and regression models to study the question. Based on the research question of interest, 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 considered four variables as alternative measures to the dependent variable, teenager mental well-being. Without better expertise in the field, we follow Orben's and consider the four choices of the response variable as part of the operational decisions we determine. The four variables are:

- "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 actuall attempt suicide?” Due to the fact that the three of the response variables are binary variables and the fourth variable can be reinterpreted as a binary variable, a logit regression model maybe approrpriate. 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 is used here to illustrate the operational decisions made without reasonable alternative choices.

I then determined two other control variables which can be include 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 variable and the independent variable.

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

4.2 Specification Curve Results and Analysis

Running the 128 models produce the following specification curve plot:

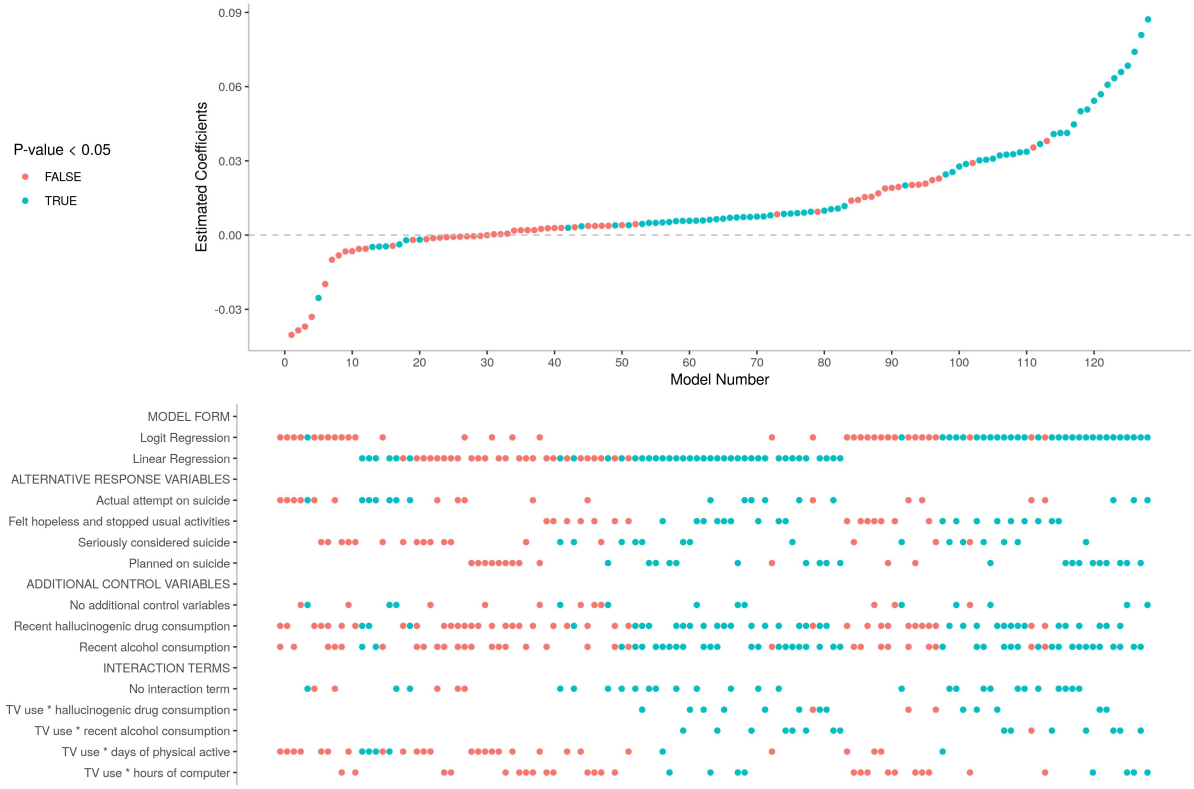


Figure 4.1: The specification curve of 128 specifications. Blue dots represent statistically significant estimates, while the red dots represent insignificant estimates

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 outstanding set of specifications in the plot. Of the models used interaction terms 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. But when interactig with days of physically active, the majority of the model estimates are not statistically significant. There are also more negative estimates. This may be suggesting that the inclusion of the two different interaction term in fact imply different underlying theories, and a closer look into the relationships between these variables is needed.

4.3 Inferential Test Results and Analysis

The last part of an SCA is an inferential test. We conducted a permutation test with 500 resampled data. We first discuss the inference on the summary statistics.

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. We reference to the permutation distribution, we found the p-values to be: 0.002 for the median, 0.146 for the proportion of positivce estimates, and 0 for the proportion of positive significant estimates. Two of the test statistics are determined to be statistically significant. This is suggesting evidence for the existence of an effect.

```
- conducted a permutation test with 500 permutated datasets
- overall results: significant median, significant prop of positive significant results
- analyzing distributions of test statistics:
    - medians: approximately normal? center at 0
    - prop of positive estimates: center at about 0.50, bumps
    - prop of positive sig estimates: largely skewed toward 0 which make sense; long tai
```

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```r
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdown))
  devtools::install_github("ismayc/thesisdown")
library(thesisdown)
```

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl.* Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime.* Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel.* Boston, MA: Wesley Addison Longman.

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`