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

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

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 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

Conducting a specification-curve analysis involves three steps: (1) Identifying the set of specifications, (2) Estimate all specifications and construct a descriptive specification curve, and (3) Conduct inferential analysis on a specification curve. This section discusses the details in each step, along with the important assumptions and concepts of the method.

1.1.1 Specifications

The first step of conducting a Specification-Curve Analysis is to enumerate the set of specifications to be considered. Before choosing our specifications, it's important to first understand 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 decisions on deciding on a specific research question/statistical hypothesis, choice of analysis method, operational decisions during the modeling process, etc. The Specification-Curve Analysis works with a specific set of specifications: the set of specifications which are (1) consistent with the underlying theory, (2) expected to be statistically valid, (3) are not redundant with other specifications in the set. The specifications used in an SCA should be valid and non-redundant as determined by the researchers working on the study. Commonly, different researchers can have disagreements over specifications. When conducting an SCA, the researchers need only to consider the set of valid specifications in their perspective. If there are lots of overlaps between the valid specifications identified by different researchers, the results of the two SCA's will be similar. If the two sets hardly or even never overlap, the results of two SCA's would expectedly be very different. As long as SCA's are applied appropriately, such a difference is likely not due to chance but may imply

something fundamentally different between the different sets of the underlying theory.

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 the 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 "appearance hair color" will be an inappropriate specification to use for conducting an SCA on this research question.

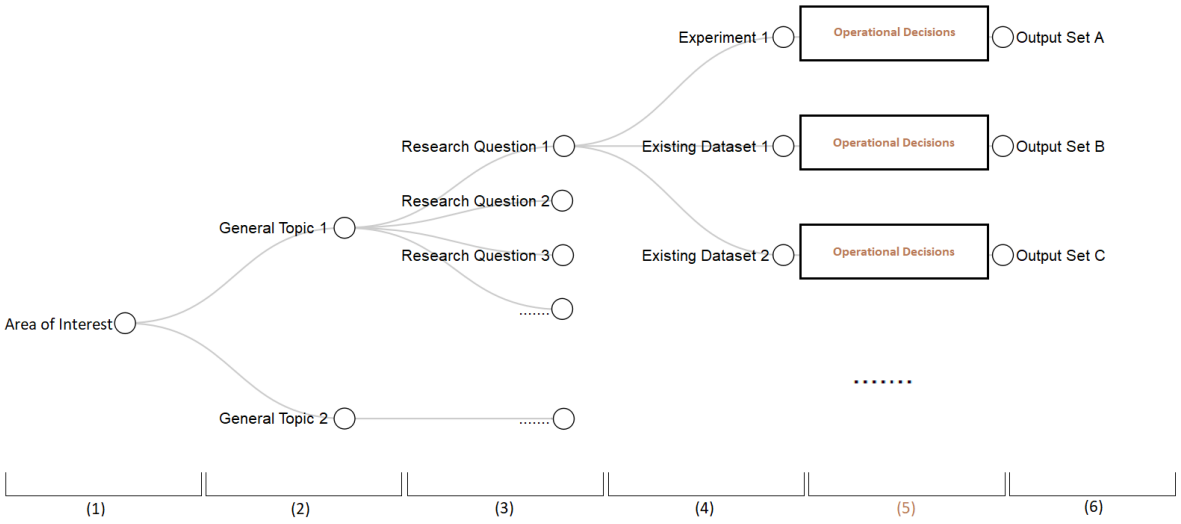

Figure 1.1: Visualization of scientific study steps

A visualization may help with understanding this idea of 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 with 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 for data collection, 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 work on modeling and analysis. If to conduct an SCA in this study, the psychologist must only consider the specific set of outputs for the study of hair color vs. teenager mental health, as specified in the following figure:

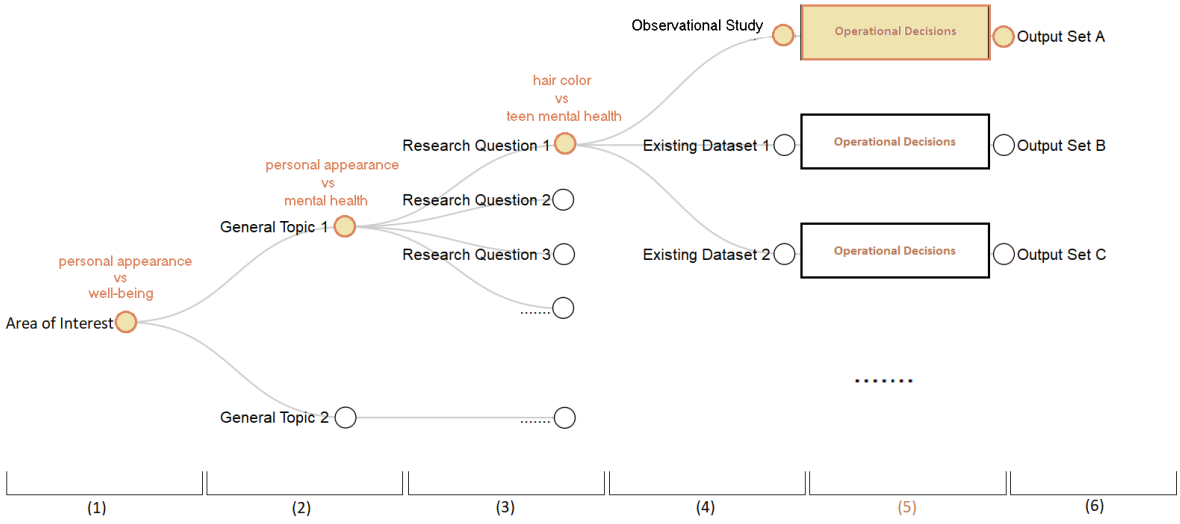

Figure 1.2: Visualization of scientific study steps

While the other decisions can be made along with an appropriate order that allows them to be connected as nodes being connected by branches of a tree, the operational decisions are more complicated. There are often different types of operational decisions,

and not all combinations of operational decisions make sense in actual modeling process. Figure 1.3 provides a visualization of the operational decisions:

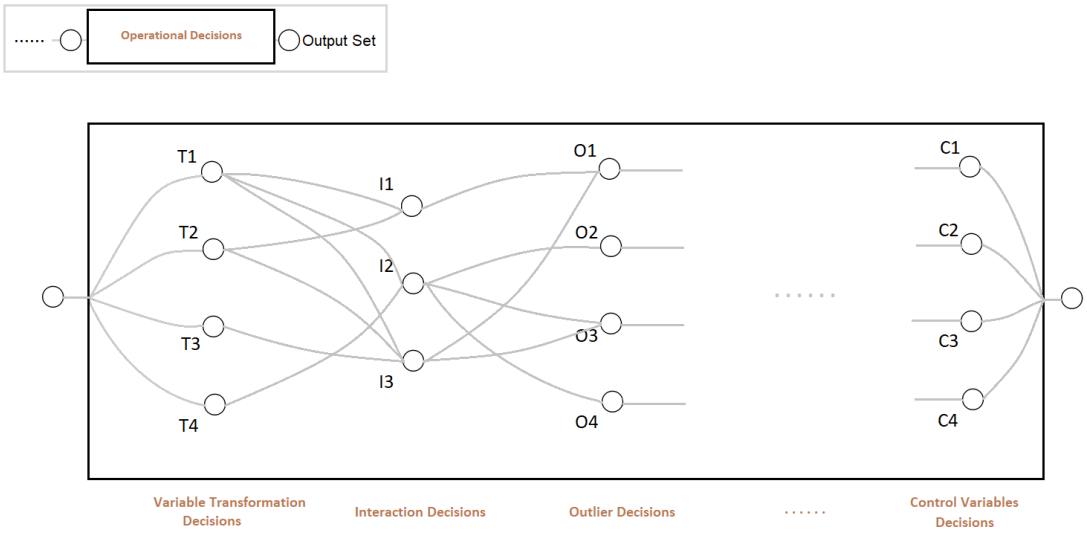

Figure 1.3: Visualization of Operational Decisions

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.

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. Thus 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 application of SCA, after enumerating the specifications of each type, all combinations of the enumerated specifications are considered, which can be visualized as in the following figure:

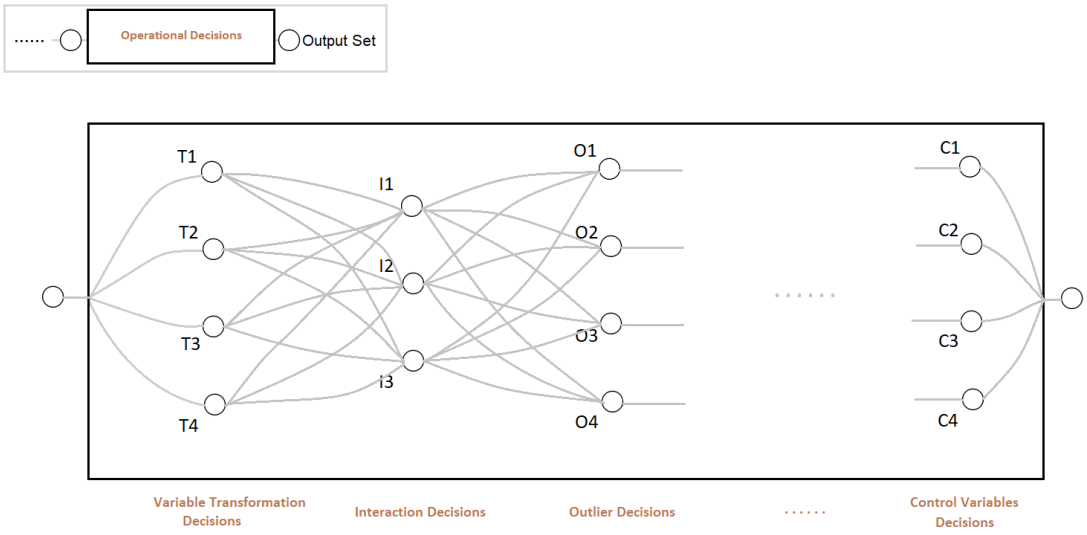

Figure 1.4: Visualization of Operational Decisions considered in existing applications

1.1.2 Specification Curve

The next step will be building a specification curve. After determining the specifications, a set of combinations of the specifications can be determined, where each combination leads to a different model to be run. Here we describe how the existing applications tend to generate combinations 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 specification 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 combinations of specifications. This makes the set of specification models to be huge and difficult to computationally work with. 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 specification 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 can be much more complicated in real life. In case of having a large number of specification models that brings computationally difficulty, a random subset of the specification models can be used instead.

Now all the models have been determined, the next step is to run all 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.

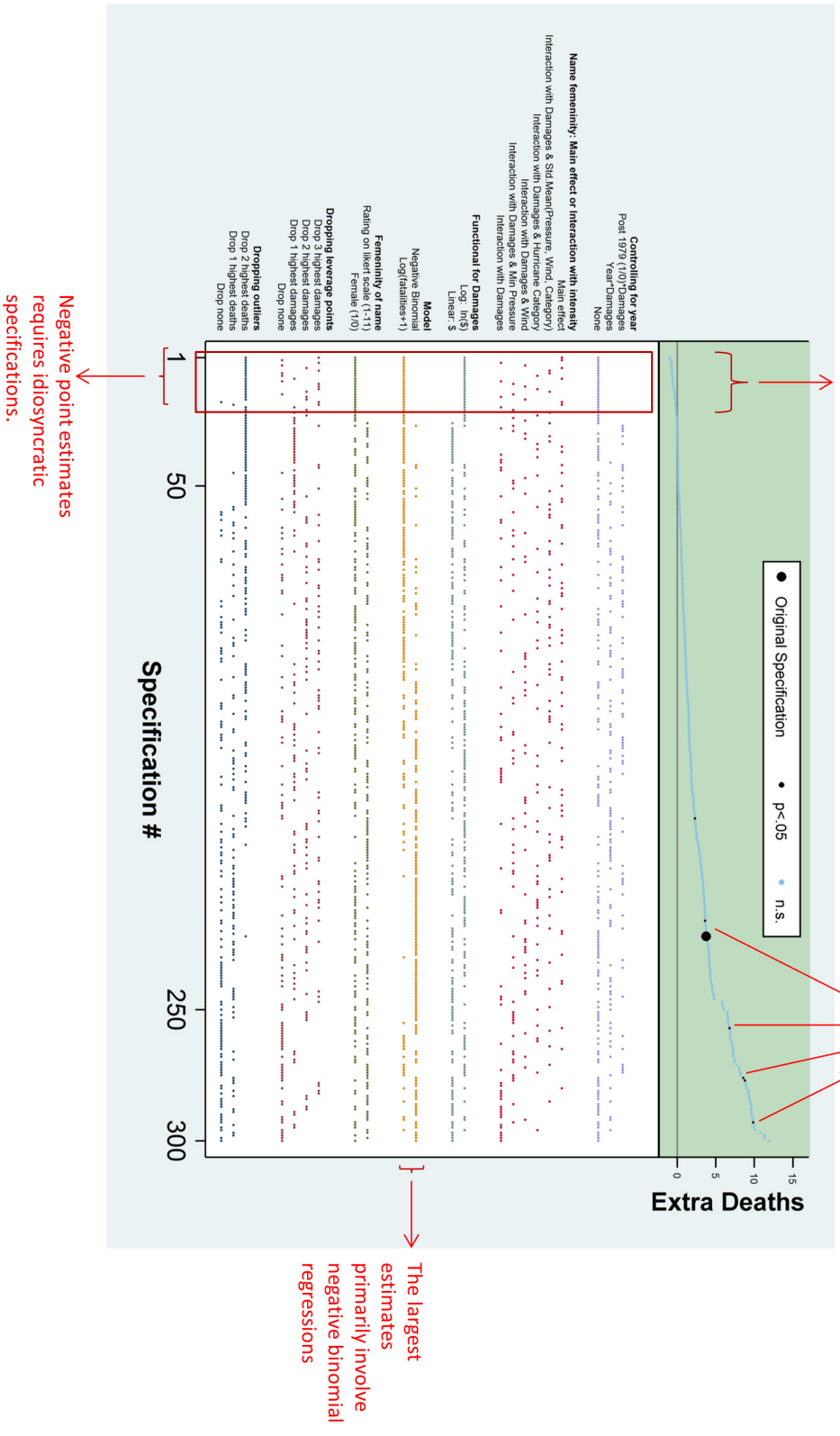

Figure 1.5: Specification Curve

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 is the curve of the estimates from each of the models, ordered from lowest value to highest value. The vertical axis is the numerical value for the estimates, and the values on the horizontal axis represent the set of specifications used for this specific model, represented by dots in the bottom plot. In the bottom half of the plot, each dot represents the usage of a specification. The vertical axis is the name of the specifications. For example, the first dot on the curve is an estimate from a model using the specifications: "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 Specification-Curve Analysis

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 the authors, is "*Considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?*". Although more formal and detailed guidance of conducting an inferential analysis for a single specification curve is desired, the authors only provided suggestions on how an inferential analysis may be performed. It was suggested that using the technique of resampling, one can generate an expected distribution of the specification curves when the null hypothesis is true. 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, and the authors did not decide on which ones to be more favored: 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. 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. This test statistic serves as a summary statistic of the entire specification curve. So instead of a "distribution of curves", the analysis works with the distribution of the test statistics from the curves. The p-value extracted, as claimed by the authors, will provide an answer to the proposed inferential question. One thing worth noting is the interpretation of the p-value. Although not specified in the main text, in the examples listed in the paper, the actual numerical value of the test statistic is not considered meaningful. For example, the authors did not use the magnitude of the median estimate for inference on effect size. The p-values are used for indicating how robust the effects are in response to changes in specifications. A low p-value indicates that the effect is robust in response to changes in specifications. This suggests that the result is inconsistent with the null hypothesis of no effect, indicates a strong sign for the existence of a statistically significant relationship. A high p-value indicates consistency with the null hypothesis of no effect, suggesting the failure to reject the hypothesis that no relationship exists.

1.2 Applications of SCA

The paper that proposed SCA is not yet published, however, 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. alternative variables representing adolescent mental well-being; 2. alternative 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. A table including all specifications determined by Orben in each dataset will be included in the Appendix. Here we provide an example: 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. 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.

Others In addition to the SCA analysis on the research question, Orben performed additional SCA analyses on 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's suggested that the small negative effect of technology use on adolescent mental well-being may be too small to warrant policy changes.

Overall, a small negative relationship between adolescent well-being and digital technology use was found, and the effect is suggested to be small enough such that no policy changes may be needed. However, several major issues exist in the application of SCA in this study, and the reliability of this result can be questionable. 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

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 with more details of 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.

Chapter 2

Replication and Evaluation

This section discusses the attempt to replicate Orben's study along with the assessment of the use of SCA in this study. We begin by introducing the three datasets used, which can all be found through public sources under permission. We then discuss in detail the attempt to replicate the study, including the obstacles to overcome during the replication process.

2.1 Data and Reprocessing

Three large-scale social datasets were used in Orben's study: Monitoring the Future (MTF) from the US, Youth Risk and Behavior Survey (YRBS) from the US, and Millennium Cohort Study (MCS) from the United Kingdom. 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.

2.1.1 YRBS

The Youth Risk Behavior Surveillance was first launched in 1990, and it's a biennial survey of adolescents that reflects a nationally representative sample of students attending secondary schools. Orben's study focused on the data collected from 2007 to 2015, and the same set of data was obtained. 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 characteristic values of the variables into corresponding 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 indexes, and the recent updates in survey questions result in differences of 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. Careful research and recoding are done to ensure the correct set of variables was used for the replication.

2.1.2 MTF

Monitoring the Future was first launched in the year of 1975, and it is an annual nationally representative survey of approximately 50,000 US adolescents in grades 8, 10 and 12. 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". Orben focused on the data collected from 2008 to 2016, which included 136,190 girls and 132,482 boys. The data are publicly accessible. In Orben's study, a merged dataset containing MTF data from 2008 to 2016 was used. While the MTF data for each year is publicly accessible, no access to a merged MTF dataset for the specified period have been found. From 2008 to 2016, the survey has been updated multiple times, along with one major change in data file format after RStudio's release in the year of 2011. Due to the frequent updates in the annual surveys and changes in data files, the variable names vary greatly among the available datasets. This brings excessive difficulties to obtain the same dataset as used in Orben's study for replication purposes.

2.1.3 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 participated children were aged between 13 and 15. The sample included 5926 girls and 5946 boys along with 10605 caregivers. The same dataset as used by Orben was obtained. Access to the data is open to the public but requires specific permission. 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 into the exact numerical indices matching with the 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 for this attempt to replicate.

2.2 Replication

After obtaining the datasets we began the replication of Orben's study. The replication 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.

2.2.1 SCA

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 10,004 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. And 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 studies YRBS and MTF. 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 a reasonable size, 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.

2.2.2 Bootstrapping test

The next part of the replication is to replicate the inference of the single 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 correctness of the inference. For now, we focus only on replicating the test and the result.

500 specification curves were conducted in Orben's study on bootstrapped samples for each of the three datasets. It was found that the test statistic for the single specification curves was statistically significant in all three cases. 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, 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 for parallel running. 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 boostrapped samples in this replication are different than the samples used in Orben's work, and this could result in a difference between the results of Orben's and this replication.

[ZJZJ Results will be added once the full implementation of the bootstrapping test implementation is finished for all three datasets. It is now completed for YRBS and MCS. MTF data still wrangling (very time consuming, progress <50% 4/5/2020).]

2.3 Evaluating Orben's work

A full 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 reliable. In this section, we talk in detail about our critiques on the usage of the SCA method in Orben's study, 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.

2.3.1 "one-to-many" mapping from scientific to statistical hypotheses

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. Intuitively, one could consider the different types of digital technologies to mean very different things. For adolescents, the different digital technologies may have very distinct functions. It would not be intuitively right to say that the functions of social media are identical to those of TV for modern teenagers. The interactive nature of certain types of digital technologies makes a distinction between them and other devices, which only transmit and output information. Intuitively, it does not sound right to consider that the distinct digital technologies would have similar effects on adolescent well-being. Therefore, it may not be appropriate to consider the different technologies as alternative representations of an integrated category. However, throughout Orben's study, a different type of digital technologies including "TV use", "Social Media use", "Time spent on electronic games" are considered alternative variables to use representing "digital technology use". With a glance, it seems like multiple related but different research questions have been collapsed into one. As discussed by Gelman and Loken (2013), the (scientific) hypotheses described here correspond to multiple *statistical hypotheses*. Whatever results conducted by this study, due to the broad research question, they would be fit into theories easily. This will result in the multiple comparison problem in the study, even if the scientists were not intended to do so.

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 (such as interactive usage of social media and passive consumption of social media) into an overarching category is suboptimal. (Bessiere, Kiesler, Kraut, & Boneva, 2008; Burke, Kraut, & Marlow, 2011; Verduyn, Ybarra, Resibois, Jonides & Kross, 2017) It's also been found with an overall reviewing on the consequences of interacting with social network sites for subjective well-being that, "passively" using social networks result in a negative relationship with subject well-being, while "actively" using social networks has a positive relationship with subject well-being. If different styles of using social network can have different relationships with subject well-being, it does not sound appropriate to consider the general usage of social media would have identical relationship with subject well-being as the usage of other technologies, such as television which provides mainly browsing of information, or electronic games which would have distinct functions depending on type of games.

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.

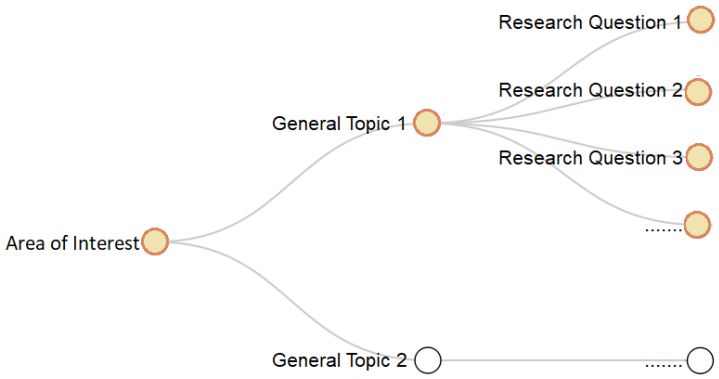

Figure 2.1: Orben may have considered multiple research questions as alternatives to each other instead of choosing a specific research question from the general topic, which is what should have been done to conduct an appropriate SCA analysis.

Such problem does not exist in the work conducted by Rohrer et al. ((Rohrer et al., 2017)) By Rohrer et al., 11 specific research questions of interest are determined, and an SCA analysis is conducted for each specific research question. 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. While there may be many different ways of measuring life satisfaction, there is much less ambiguity of what is represented by this term. Life satisfaction can be considered as being part of the category of personality, while technology use is more like a general category that is on the same level of personality.

2.3.2 Choice of Specifications

The specifications determined by Orben, due to the choice of multiple different 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.

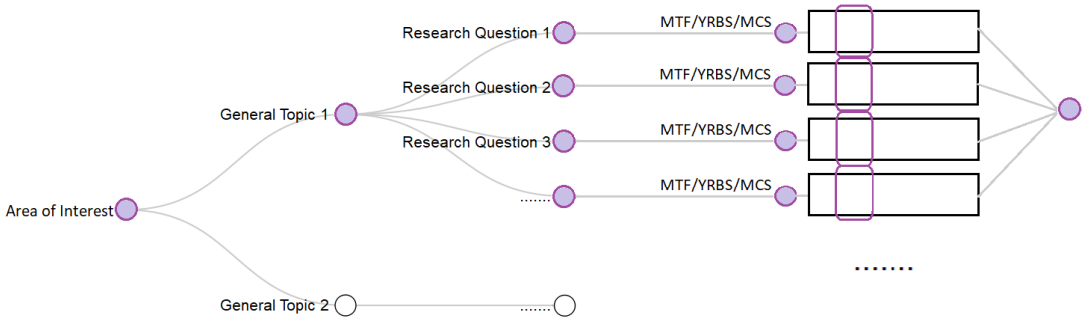

Figure 2.2: The set of specifications considered by Orben

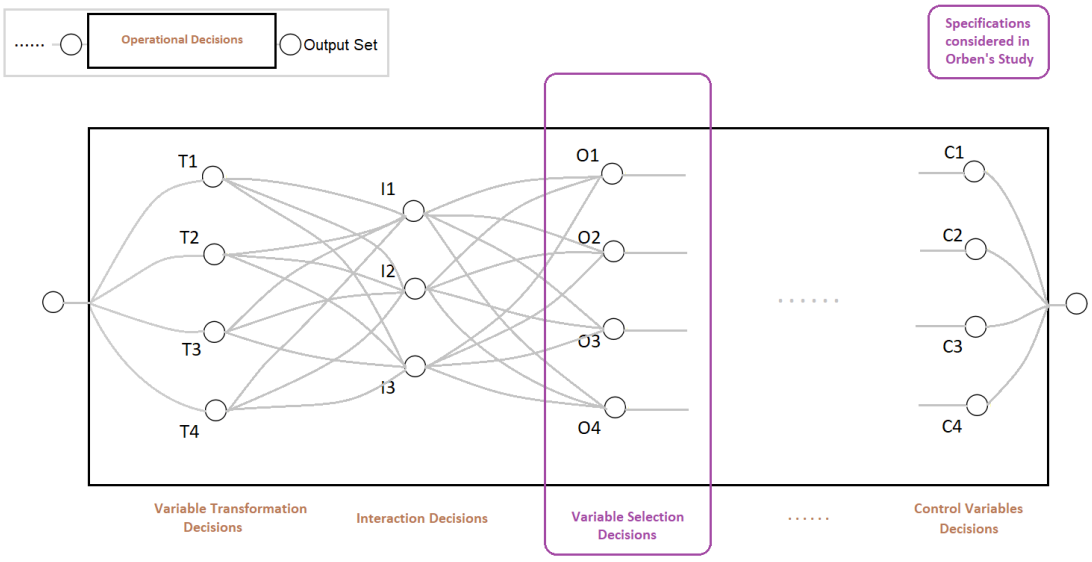

Figure 2.3: The set of specifications considered by Orben

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 an appropriate design of SCA analysis.

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

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.

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.

2.3.3 SCA interpretation

The last major problem of this study is the way Orben interprets the single SCA result. As discussed in the previous chapter, the single SCA generated is used for a descriptive curve that can provide information on whether or not the relationship appears to be happening by chance, and if a certain pattern of a true relationship is observed, if the relationship appears to be robust in response to changes in specifications. The single specification curve should not be used for any interpretation of the actual magnitude of the numerical values of the estimates. However in Orben's study, when interpreting the single generated 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. Nowhere suggested by Simonsohn et al describes this interpretation of the numerical result from a specification curve. The three examples provided in the original paper describing the method do not make such conclusions from a single specification curve but only used it to assess if the relationship seems strong and which specifications appear to have the largest effect on the estimate. When conducting inferential tests for the three examples, the medians were only used to check 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. While it is unknown of the rationale behind Orben's innovative steps when conducting the SCA analysis, Chapter 3 provides a detailed discussion of why these innovations may in fact provide unreliable results and interpratations.

Chapter 3

Inference for SCA

This chapter focuses on the statistical inference on a specification curve. Simonsohn et al. provided guidelines and suggestions on conducting an inferential test on a specification curve using the bootstrapping or the permutation technique along with three choices of test statistics. However, such process is not statistically formalized. In the followings sections, we attempt to statistically formalize the proposed inferential test for a specification curve, from its hypothesis, test statistics to its conclusion and interpretation. We also evaluate the possibility of additional inferences based on the formalized test, such as interpretation on numerical values of the test statistics, additional inference on "Dominant sign", and additional inference on the outstanding combinations of specifications.

3.1 Formalizing suggestions from paper

We start by attempting to statistically formalize the inferential test on a specification curve. Simonsohn proposed the inferential test in the structure of a hypothesis testing, including three main parts: statistical hypothesis, test statistics and null distribution, and inference on the test statistics. In the following sections, we discuss and evaluate each of the three parts proposed, and provide formalizations of each part.

3.1.1 Statistical Hypothesis

The paper indicated that the inferential test is designed to answer the question, "*Considering the full set of reasonable specifications jointly, how inconsistent are the results with the null hypothesis of no effect?*" The authors never explicitly limits the type of research questions that can be studied by an SCA to be causal research questions only. The examples provided in the paper all worked with correlation problems, instead of causal problems. While the authors indicated the null hypothesis to be "no effect" here, the true null hypothesis they meant should be "no relationship/correlation" or "no effect". In the case of the model form being fixed to linear regressions, the null hypothesis can be rephrased simply as $\beta = 0$, regardless of the type of relationship being studied. It is then natural to rephrase the alternative hypothesis as $\beta \neq 0$,

"exists a relationship/correlation" or "has effect".

3.1.2 Test Statistics and Null Distribution

The next part of a hypothesis testing is to determine the test statistics, the null distribution, and fin the p-value for inference. Simonsohn proposed three choices of test statistics: 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 ($p < 0.05$). Due to the great flexibility of choosing specifications, it is difficult to determine the null distribution of these test statistics analyically. Simonsohn suggested using the resampling technique to find the expected distribution of specifications curves when the null hypothesis is true: permutation technique for data with random assignment, and bootstrapping technique for data without it. This is a reasonable approach for estimating the null distribution of specifications curves, and thus determine a null distribution of the test statistics. But this approach can be computationally expensive. To generate a distribution, a large number of specification curves based on resampled data will be needed. Generating a single specification curve can already be computationall expensive, especially when the number of alternative specifications is large and/or the model form is complicated. Generating, say, several hundreds of such specifications curves can take several days or even weeks to finish, even when running in parallel on a powerful server. This brings great difficulty for conducting an inferential test for SCA in real life. The process could be well simplified if a reference null distribution can be more easily generated using statistical theories. For each of the three test statistics, we attempt to provide alternative ways of constructing the null distribution of test statistics, without having to generate the full expected null distribution of specification curves, using statistical theories.

1.

2.

3.

3.1.3 Conclusion and Interpretation

The final step of conducting a hypothesis test is drawing conclusions from the p-value. The way of interpreting a p-value in this case is same as interpreting a p-value in other cases. With a significance level α chosen, if $p < \alpha$, then we reject the null hypothesis. Otherwise we fail to reject the null hypothesis. In this case, the p-value may only used for assessing if and how the results from the single specification curve are consistent with the null hypothesis of no effect/no relationship. It is important to note that this inferencial test on a specification curve is essentially a hypothesis testing, and the conclusions that could be drawn from a hypothesis testing is fixed: whether or not we reject the null hypothesis. It is dangerous to make additional conclusions despite the whole test procedure is designed to be a simple hypothesis test. Orben, for example,

drawn conclusions on the numerical values of one of the three test statistics, alongside with the normal interpretation of the p-value. No justifications have been provided for the additional conclusions being drawn. In the later section, we will discuss in more details of why this additional interpretation of test statistics can be misleading.

3.1.4 Formalized Procedure

We can now formalize the procedure of conducting a inferential test on a single specification curve as the following:

1. Indicating the hypotheses: H_0 : No effect/No relationship; H_1 : Has effect/Exist relationship.
2. Generating test statistic(s):
   (a) the median overall point estimate from the specification curve
   (b) the share of estimates in specification curve that are of the dominant sign
   (c) the share that is of the dominant sign and also statistically significant ($p < 0.05$)
3. Compute null distributions and determine p-value(s):
4. Drawn conclusion: if $p > \alpha$, failed to reject null with statistical significance α. Otherwise, reject null with statistical significance α.

3.2 Additional Inference

A single specification curve provides information on how robust a model appears to be in response to changes in specifications. The inferential test on a specification curve answers the question of whether there is evidence of an effect/a relationship. Could we learn more from an SCA? Orben attempted to interpret the numerical value of the median overall point estimate from the specification curve as representing the magnitude of an effect/relationship, is this an appropriate inference? In the following sections, we discuss and evaluate some possible additional inference, existing in the current applications or not, on an SCA.

3.2.1 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 this case, however, every point estimate in a specification curve would have been a reasonble estimate of the effect/relationship. These estimates are likely not all close to each other, they 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 of the existence of an effect/a relationship, the proposed SCA 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, it is questionable if the different point estimates based on different set of specifications can be considered as the estimates of the exact same thing. Certain choices of specifications may lead to differences in the exact model forms, and the numerical value of different point estimates may not mean the same thing. 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 estiamted effect/relationship. With the great flexibility of choosing specifications in SCA, it is highly 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 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.

While it is difficult and nearly unrealistic to measure magnitude of effect/relationship based on summary statistic of the point estimates, we can still draw conclusions about the estimated sign of the effect/relationship using the dominant sign, the sign of the majority of estimates. It may not be appropriate to directly draw conclusion based on the dominant sign. If 51% of the estimates are positive and 49% of the estimates are negative, the dominant sign is positive. But with the proportions being approximately 50%, claiming that the estimated sign of the effect/relationship is positive seems inappriate. It is intuitively right that if the dominant sign shares a proportion that's large enough, we can claim that the overall estimated sign of the effect/relationship is of the dominant sign–but how large would be large enough? We would need to setup some threshold.

3.2.2 Inference on "Dominant sign"

- How "dominant" a sign is can we draw conclusion of pos/neg/neutral effect/relationship? (set up threshold)

3.2.3 Outstanding combinations of specifications

- set of specifications which generate outstanding estimates, differing from the majority of the estimates. (In terms of signs)
- If an outstanding combinatinos of specifications, could be potentially suggesting different underlying theories.

Critiques of SCA: change in control variables could be potentially suggesting different theories.

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

```
# 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