“Indexes and Multiple Hypothesis Testing”

TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN ECONOMÍA

Alumno: Marcelo Gómez R.
Profesor Guía: Damian Clarke

Santiago, junio de 2024

Indexes and Multiple Hypothesis Testing

Marcelo Gómez R.*

June 21, 2024

Abstract

Multiple hypothesis testing issues have appeared in the economics field over the last decade, providing a broad palette of methods designed to address this problem. The social sciences, in general, have greatly benefited from these advancements. Issues related to testing multiple hypotheses with a single treatment variable have been growing in economics over the last decades. However, other methods for constructing indices were in use before Anderson's. Until today, researchers have not provided, or at least discussed, a structured set of ground rules to properly use these methods. In this thesis, we generate a statistical framework, primarily in the context of program evaluation, to assess the performance of different indexing techniques currently employed in the social sciences literature. Specifically, we evaluate the index proposed by Anderson (2008), the index produced from Principal Component Analysis and finally, a simple sum of the standardized variables index. We find that the way such indices are generated can lead to important differences in decisions related to rejecting or not rejecting null hypotheses of the significance of grouped outcomes.

JEL codes: C12, D04, C43
Keywords: Multiple Hypotheses, Index, Principal Component, Anderson

*University of Chile. Email margora@fen.uchile.cl.

1 Introduction

Statistical inference is critical to all kind of experimental researchers. Nowadays, we may say that the quintessential for economic analysis is the instrument called *Hypothesis Test* (HT). In this context, imagine that you want to assess a training program impact. To do this, we can hypothesize that it will have a positive impact on worker's technical abilities. To prove our hypothesis, the customary procedure is to define a null hypothesis H_0 (stating that the training program will not affect workers' technical abilities) and an alternative hypothesis H_1 (indicating that it will indeed have an impact on workers abilities). Therefore, if the null hypothesis gets rejected (when it is actually false), it *should* provide evidence to support our statement regarding the causal effect within our experiment. Nevertheless, there still exists the possibility of some error being committed in our procedure. There are two errors that appears from the latter insight. The likelihood of erroneously reject H_0 when its true (called Type I error), or not reject H_0 when its false (called Type II error). Ideally it will be better if both Type I and II errors does not occur.

For the last example, we implicitly assume that just one hypothesis was studied. But what would happen if we conducted an experiment involving multiple hypotheses? Here is where the concept of *Multiple Hypotheses Testing* comes into play (herein called MHT). When conducting MHT, it is crucial to be careful to avoid falsely rejecting too many null hypotheses. Furthermore, if we reject null hypotheses without taking into account the fact that multiple tests were performed, the ideal correction of errors will not be achieved. Now, from a formal statistical point of view there are two possibilities to deal with multiple comparisons, which have been widely studied by statisticians. We can either adjust p-values while keeping the significance level unchanged or, conversely, adjust the pre-specified significance level while keeping p-values the same.[1]

In this thesis, our aim is to develop a practical framework that sheds lights upon indexing and addresses the common concerns associated with MHT. Essentially, our goal is to produce a guide that outlines key points about program evaluation using indexes from an empirical standpoint, with a primary focus on conducting simulations and using adequate real-life data used in others papers.

The usual approach in the generality of papers is that they use different indexing techniques without any prior rules to determine which method for creating an index is *better* for the particular study. Consequently, the question of which index should be implemented has been left entirely to the discretion of researchers. Recently, it has become more common to use MHT in economic literature. The reason is that

[1] Others variations have also been developed, such as Bayesian or quasi-Bayesian methods, to account for MHT.

it allows researchers to distinguish significant coefficients that may appears by *chance*, i.e., regardless of the presence or absence of treatment effects.[2]

At present, there are specific ways to conduct valid inference for cases with multiple hypotheses. Two important approaches are Generalized Error Rate (GER) Correction Strategies (Lehmann and Romano, 2005) and Dimension Reduction (DR) techniques using indexes. The latter is our main focus of study.[3] There are two usual practices to control error rates: *Family-Wise Error Rate* (FWER) and *False Discovery Rate* (FDR).[4] Finally, the difference between GER and DR is that the former consider the total number of hypotheses conducted by the researcher, but reports the $p-$values based on the number of reported results. On the other hand, DR consider a reduction on the number of hypotheses to take into account MHT.

Another direction that the literature has taken is to assess the need for adjustment using a game between two agents: a social planner (who could be a policymaker or a academic journal editor) and a researcher, where a economic model with incentives is developed (Sterling, 1959; Viviano et al., 2023; Tetenov, 2016). Particularly, in economics and social sciences, there is a focus on positive results that affects topics like transparency and reproducibility in empirical research (Yong, 2012). Hence, these aspects have been receiving renewed attention in recent decades.[5] As Wasserstein and Lazar (2016); Greenland et al. (2016) indicates, some erroneous conventions about $p-$values are related to statistical misinterpretation in the scientific discipline. Performing multiple tests and only presenting the p-values that yield the desired outcome is considered tremendously problematic. Furthermore, excessive *exploratory analysis* can lead to inappropriate practices. Nevertheless, the use of multiple test procedures cannot protect against the bias caused by data fishing (Bender R, 2001). Given the above, some journals and researchers advocate guidelines for authors and reviewers that performs econometric analysis (Simmons et al., 2011; Miguel et al., 2014). The downside is that some of these practices increase the odds of Type I errors that may not be present in other samples, but equally problematic in surveys.

[2]These procedures have been in use since the 1950s, and they are even mandatory for some journals or institutions (Food and Administration, 2022).

[3]There is a third way of account for multiple hypothesis is through Machine Learning techniques that reframe the problem of estimation as one of *prediction*. This method is rather new in economics and discussions on the econometrics context can be found in (Mullainathan and Spiess, 2017).

[4]For GER methods, there is vast literature discussing ways to adjust inference, aiming to reduce the general rate of false positives (Bonferroni, 1935; Holm, 1979; Westfall and Young, 1993; Romano and Wolf, 2005; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001).

[5]Discussions of related topics such as transparency, reproducibility and credibility in economics can be found in (Christensen and Miguel (2018); Miguel (2021)). Others practical problems related to hypothesis testing are Cherry-Picking, $p-$hacking, Harking and data-snooping or data-dredging. Explanations and examples in the field of Biostatistics are presented in (Andrade, 2021).

There is a broad range of experiments where DR techniques can be implemented. The customary procedure includes transforming outcomes into indexes since researchers are often allowed to extract a lot of outcomes of interest (for example, see Cohen et al. (2023); Chong and Valdivia (2023)). [6] In Political Science, (Denny et al., 2023) explored how extortion can alter political preferences using ICW and a *simple* Average (SA) to construct the indexes.[7] Data DR techniques have been in use for a while in economics. For example, Principal Component Analysis is broadly used to measure *wealth*, researchers argue that this method is employed due to the difficulties in obtaining accurate income data (e.g., (Vyas and Kumaranayake, 2006; Houweling et al., 2003)). For Index tests, the two more commonly used methods are the index generated using ICW, as provided by Anderson (2008), and the index generated from Principal component Analysis[8]. In addition, we will use the simplest index as baseline. This index can be created by summing of standardized variables of interest (hereinafter *Summary index* or SI).

Finally, our focus on indexing methodologies arises from the dramatic use of MHT corrections in the field of economics, specially in some sub-fields such as evaluation program, socioeconomic studies, etc. (Viviano et al., 2023) reports that, since 2014, the use of MHT corrections increased from 0 to 39% in "top 5" economic journals, with 54% of this increment involving index adjustment[9]. Aditionally, Figure A1 illustrates the trend in four top journals between 1958 and 2008.

The thesis is organized as follows. Section 2 delves in Multiple Hypothesis Testing, offering a review of literature related to controlling FWER and FDR, which will be later linked to results in Section 4. Additionally, provides a discussion not attained in the literature about the relationship between Index and MHT. Section 3 explains the different index tests, their procedures, and includes some practical comments involving MHT. Section 4 presents Monte Carlo simulations to assess indexes performance in various scenarios, particularly providing a comparative measure for FWER and the power of the index tests in each scenario. Section 5 describes an empirical experiment using papers related to program evaluation. Finally, Section 6 provides insightful discussions about practical concerns regarding indexes.

[6] For instance, in psychology, Baghumyan (2023) conducted a experiment about discrimination using only an index generated with the ICW method. On the other hand, Evans et al. (2023) evaluated child development in Mexico, using indexes generated through Principal Component Analysis (PCA), Inverse Covariance Weighting (ICW) and the Sum of Standardized (SS) variables.

[7] There are some specific branches that includes Factor Analysis in the index construction analyses. However, given that the spirit of Factor Analysis (and variations, for example, in macroeconomics the name Dynamic Factor Analysis is not uncommon) is somewhat upside down compared to indexes that do not rely on a structural model. That is why FA is not a research topic in this thesis.

[8] For an introduction to this topic, see Hotelling (1933); Anderson (1963).

[9] Allee et al. (2022) found 95 studies that implemented Principal Component Analysis in the accounting research field.

2 From Hypothesis Testing to Multiple Hypothesis Testing

The following section introduces the concept of Multiple Hypothesis testing. This discussion aims to achieve two goals. First, it presents the most well-known method to correct for problems of multiple hypotheses, called the FWER correction, and demonstrates that can be used as benchmark for our simulations with index tests. Second, it develop an argument to highlight important questions that researchers are not asking in the current literature, which are potentially relevant to the investigation results and their consequences.

2.1 Motivation

Consider a null hypothesis H_0 and suppose we have data X with distribution $\mathbb{P}$ is available to deal with the following problem: given an observation $x \in X$, how do we decide whether to accept or reject a hypothesis about $\mathbb{P}$? In a simple HT, there are four possible scenarios summarized in Table 1. To minimize the probability of falsely rejecting H_0 when it is true, we typically choose a low and *arbitrary* value for the odds of making a Type I error[10].

Table 1: Possible scenarios associated with simple Hypothesis Test

		Null Hypothesis (H_0) is:	
		True	False
Decision upon H_0:	Reject	Type I error, $Pr = \alpha$	Correct, $Pr = 1 - \beta$
	No Reject	Correct, $Pr = 1 - \alpha$	Type II error, $Pr = \beta$

Note: There are two kinds of errors when HT are performed. Type I and II errors, their probabilities are α and β, respectively.

Now, the idea presented above does not take into account the Type II error. Therefore, it does not provide any correction for β. As a result, the power of a test becomes crucial, defined as the probability that the null hypothesis is correctly rejected when the alternative hypothesis is true, denoted by $1 - \beta$.

A Type I error occurs whenever we test an hypothesis on an outcome. Consider the case where we are solely interested in controlling the Type I error. In such a scenario, if a test over a -it may be said-

[10]Called size of a test, α, where α takes on conventional values such as $\{0.01, 0.05, 0.1\}$

second outcome is performed, there exists the possibility of committing none, one or two Type I errors. To generalize, we can calculate the odds of incurring *at least* one Type I error. The easiest way to do this is by assuming that each hypothesis test is independent of the others. Thus, the last-mentioned probability can be obtained from:

$$P\left(V \geq 1\right) = 1 - P(V = 0) \geq 1 - (1 - \alpha)^M > \alpha \tag{1}$$

Where $M \in \{2, 3, \dots\}$ is the number of variables being tested and V is the number of Type I errors made.[11] Evidently, the probability increases with the value of M. For that reason, we encounter a different problem than in simple HT. [12] A multiple testing framework, similar to the simplest case, can be classified based on the significance of a test. In this manners, *discoveries* can be defined as true or false. Refer to Table 2 to formalize the idea[13]:

Table 2: Outcomes when testing M hypotheses

	Null Hypothesis (H_0) is true	**Alternative Hypothesis (H_1 is true)**	Total
Test is declared significant	V	S	R
Test is declared non significant	U	T	$M - R$
Total	M_0	$M - M_0$	M

Note: For a given hypothesis, the null could be true or false, and the corresponding test could reject or fail to reject the null. Notation follows V as the number of false positives (Type I error or *False Discoveries*), S is the number of true positives (*True Discoveries*), T is the number of false negatives (Type II error), U is the number of true negatives and $R = V + S$: Number of rejected null hypotheses (*Discoveries*).

Multiple hypothesis testing has been widely studied, leading to the development of numerous procedures to alleviate its main problem. Indexing techniques have arisen as a solution to the same problem

[11] A more detailed way (without assuming independence) of seeing this formula is $P(V \geq 1)$:

$$\begin{aligned} P(V \geq 1) &= 1 - P(V = 0) \\ &= 1 - P(\text{do not erroneously reject any null hypotheses}) \\ &= 1 - \left(\bigcup_{j=1}^{M} \{\text{do not erroneously reject } H_{0j}\} \right). \end{aligned}$$

[12] To illustrate this, we calculate the probability obtained from Equation 1 for various significance levels and different number of hypotheses tested. The results are presented in Figure A2. Note that when the number of hypotheses increases, it raise the probability of make at least one Type I error, sooner or later depending on the significance level predefined.

[13] Note that is not a formal framework for the MHT issue, before formalize it, we will discuss some insights about MHT and inference, taking somewhat philosophical approach.

(though not as generalized like MHT) but from a different framework. As of today, there is no a clear connection between indexing and MHT in terms of comparability, despite both methods addressing the same problem.

In this context, the literature almost does not provides profound discussions about indexing and neither about its relationship with MHT. Thus, the next subsection opens up an insightful discussion about how the creation of indexes could affect the research goals.

2.2 When to (and not to) adjust?

The introduction of MHT methods corrects -at some level- latent issues in research evaluation methodologies given that it provides a framework to properly control GERs to avoid erroneous conclusions, and hopefully trying to make the right choice as often as possible, i.e., diminishing the probability of making a Type II error.

Then, the first natural question is, when do we encounter a MHT problem? List et al. (2016) outlines the prevalent cases in which MHT should be considered:

> *"In this setting, different null hypotheses arise naturally for at least three different reasons: when there are multiple outcomes of interest and it is desired to determine on which of these outcomes a treatment has an effect; when the effect of a treatment may be heterogeneous in that it varies across subgroups defined by observed characteristics (e.g., gender or age) and it is desired to determine for which of these subgroups a treatment has an effect; and finally when there are multiple treatments of interest and it is desired to determine which treatments have an effect relative to either the control or relative to each of the other treatments."*

Now, an important issue that emerges is whether we actually need to adjust our experiment. There is no direct consensus about which experiments need to be adjusted, but there are some insights that can help shed light on the issue.

For a basic example, consider a 2×2 framework where we have an *old* and *new* teaching method, students' gender (male, female) and some measure of student academic achievement. The study's objective is to approve the new teaching method's effectiveness. In this context, we want to assess whether the teaching method exhibit an effect or not. It is important to note that other contrasts (or groupings)

are primary descriptive or mostly independent of the main objective. Therefore, any further investigation into gender effects interacted with teaching methods becomes secondary. Consequently, in this case, multiplicity adjustments have arguably become unnecessary since a harmful erroneous discovery is restricted to the teaching method effect (Frane, 2015). Furthermore, it is worth noting that there are authors who criticize the use of MHT and presents arguments against correcting hypothesis testing with MHT procedures (e.g. Rothman (1990)). Streiner (2015) discusses both viewpoints, those in favor and against it, highlighting compelling issues in both positions.

To understand when and how MHT should be used appropriately, one option is consider the intended audience for the experiment's results, such as policymakers or journal editors. Another idea is to clarify grouping in empirical research is to provide a pre-specified analysis plan.[14][15] Now, the decision of including MHT should be considered carefully. For the first case (pre-analysis plan) it *may* even be mandatory. Recently, PAP have gained traction in empirical microeconomics. Researchers in this field publish their hypotheses and methods for dealing with multiple hypotheses (if necessary) on recognized public platforms like the AEA RCT Registry, prior to collecting the data. This approach aims to increase the power of their studies (Anderson and Magruder, 2022).

The next relevant question is, how indexing relate to MHT? A slightly different framework for multiple comparison involves making *overall* decisions regarding a set of outcomes. As explained by Viviano et al. (2023), a policymaker might need to decide whether to implement a reform based on its effects on both education and health, for example.[16] In this case, it is optimal to report a single test based on an index created from the outcomes. Consequently, in the case of indexing, we depart from specific MHT procedures, and the question of adjustment for multiple comparisons *does not arise* (or is significantly reduced). This simplification reduces the dimension of the problem to that of the simple HTs.

Index test seems to offer an appealing solution for situations like the one mentioned above. Nevertheless, adopting this approach raises another kind of questions. For instance, let's consider the scenario where we create an index from five outcomes and our goal is to determine if a treatment variable had an effect on the index. The customary procedure would be to conduct a hypothesis test of the treatment on

[14] For guidance on creating a pre-analysis plan in economics, refer to Olken (2015).

[15] Anyway, the latter approaches reveal two distinct types of data. Data that its actually gathered with a specified plan before any assessment, such as Randomized Controlled Trials (RCTs). The second type of data that it is collected without prior hypotheses in mind. Nevertheless, regardless the researcher's intentions in either case are *confirmatory* or *exploratory*, a distinction that can be blurry in practice.

[16] However, this may not apply if policymakers are interested in only one of the outcomes or more than one independently. In such cases, other MHT procedures not involving indexing may be more appropriate.

the index. Now, if we find that the effect is statistically significant, what does this means? Does it imply that all the variables were affected by the treatment or, just one, two or maybe three out of the total were affected? Currently, the literature on indexing has not adequately addressed this issue. The use of index techniques is generally employed with little more justification than to correct for MHT.

Let's think again about the teaching example. Suppose the treatment variable represents the students who received the new teaching method, also, we have five different measure of academic achievement. We then create an index from these five variables. Is the statistical performance of the index the same if we are interested in finding at least one significant effect as if we want that at least half of the variables to show some effect from the treatment? What would happen if all (or none of) the variables had an effect from treatment? In Section 4, we address these questions.

Finally, there are not conclusive statements about how and when to adjust, rather than insights for specific situations as the aforementioned (Streiner and Norman, 2011). Even though this section discusses the problem in an incentive/pre-conception framework, it leaves aside specific statistical limitations that could come to light from the MHT procedures revised later. That is why, an important point to note in classical hypothesis testing is the inherent subjectivity in the interpretation of data (as somewhat discussed in the insights from this section) to provide a more accurate and flexible statistical assessment (Berger and Berry, 1988).

2.3 Multiple testing and how to address the issue

Let us formalize, following the notation from Lehmann and Romano (2022), the generalization for simultaneously testing $M \in \mathbb{N}$ hypotheses H_m $(m = 1, \ldots, M)$. Suppose data X is available from some model $\mathbb{P} \in \Omega$. The multiple testing problem can be denoted as:

$$H_m : \theta \in w_m \quad i = 1, \ldots, M \tag{2}$$

Given w as a subset of Ω. Let $T = T(\theta)$ denote the set of true null hypotheses when θ is the true probability distribution. Then, $m \in T(\theta)$ if and only if $\theta \in w_m$.

The literature has been primarily focused on two approaches: *Familywise Error Rate (FWER)*, which studies the probability of rejecting at least one null hypothesis in a family where the null hypotheses are

true.[17] This approach jointly consider all hypotheses as a family in order to control the odds of making at least one Type I error at a level α.[18]

$$FWER = P(\text{reject any } H_m \text{ with } m \in T(\theta)) \leq \alpha \quad \forall \theta \in \Omega \tag{3}$$

Second, the *False Discovery Rate (FDR)* (proposed by (Benjamini and Hochberg, 1995)), is defined as the expected proportion of errors among the rejected hypotheses. Let the *False Discovery Proportion* (FDP) be defined as $FDP = V/R \cdot D\{V > 0\}$, where $D\{\cdot\}$ denotes the indicator function. Then, the FDR is given by the expected value of the FDP and the goal is to control it to be a small proportion for some $\gamma \in [0, 1)$.

$$FDR = E(FDP) \leq \gamma \quad \text{for any } \theta \tag{4}$$

Table A1 indicates methods developed for both approaches[19]. Nonetheless, FWER literature comes with criticism. It is been proved that diminishing the Type I Error rate provokes greater Type II errors rate (Smith et al., 2006), and it is known that the FWER is too conservative and negatively affects the power of tests. Furthermore, the FWER has been criticised as too conservative in cases when too many multiple hypotheses are tested at the same time (Chen et al., 2017).

The methods presented in this final subsection consider essentially all the hypotheses that the researcher is interested in testing. Therefore, it produces $p-$values for every hypothesis tested, this could be M or subset of hypotheses. Nonetheless, one goal of DR methods is to shrink the number of hypotheses to the minimum. As mentioned before, GER methods are best suited for questions related to independently assess the statistical significance of variables of interest. Principally, they are applied individually to each variable in a variety of hypotheses.

Finally, as mentioned earlier, FWER aims to control the probability of committing at least one Type

[17] We say that FWER is weakly controlled at level α. A different approach is to strongly control FWER at level α, i.e., if $FWER \leq \alpha$ for possible constellations of T and T^C, where T^C denotes the set of false null hypotheses when θ is the true probability distribution.

[18] This proposition can be generalized to the probability of k or more false discoveries. In the literature this is denoted by *k-Familywise Error Rate* (Lehmann and Romano, 2022).

[19] Exists a third method called *positive FDR* (pFDR), introduced by Storey (2003). Later research in this setting can be found in Goeman and Solari (2011). Fourth, *The Closure Method* (introduced by Marcus et al. (1976)) that rejects a intersection of hypotheses if and only if every joint hypothesis that contains the intersection is rejected. Some procedures are Simes's identity, Hommel's method and other classical joint tests like Fisher's F test. More details about this method can be found in Lehmann and Romano (2022). Finally, Bayesian (and quasi-Bayesian) approaches as well have been developed, introduction and explanations can be found in Efron et al. (2001); Berry and Hochberg (1999). Explanations, advantages and caveats of the main methods use in each approach can be found in Goeman and Solari (2014).

I error. In Section 4, where we present statistical simulations, we will introduce some scenarios with only one variable providing a true effect, i.e., rejecting the null hypothesis for the effect of this variable (on a dependent variable, as will be explained in detail later) would be the lower bound of the FWER's probability. Then, the proportion of rejected tests will be our simple comparative measure between MHT and index tests.

3 Index Tests

Indexing methods, as mentioned before, provide a solution to avoid the use of MHT techniques. Their goal is to reduce the probability of incurring in false positives, following the principles of FWER. The fundamental principle of indexing techniques involves dimensional reduction of data. This process consists in finding an index that summarizes all the information of interest, effectively reducing N outcomes to one outcome that contains the majority of the important information. But, ¿how do you decide which outcome is more or less relevant? In this section, we explain how the indexes generated from Anderson, Principal Components and Summary methods proceed to create an index. These indexes rely on different principles to weight and aggregate outcomes, and those methods are summarized in Table 3.

Table 3: Indexing corrections

Indexes	
Methods	Comments
Summary index	Simple sum of the variables
(Hotelling, 1933)	Principal Component weighting
(Anderson, 2008)	Inverse Covariance weighting

Finally, consider that the primary focus of this study is on the use of Index as dependent variables in estimation. Therefore, we have excluded the analysis of Index tests when used as a regressors or independent variables. The reason for this exclusion is that when using it as the dependent variable, philosophical issues as the ones mentioned in Subsection 2.2, arise. Contrarily, when the index is used as a regressor, probably the most convenient way to manage MHT problems is through GER techniques.

3.1 Methods

Let's formalize, consider $Y_i \in \{AI_i, PCI_i, SI_i\}$ which is an index formed from variables y_i^m where $m \in \{1, \ldots, M\}$ indexes each variable of interest and $i \in \{1, \ldots, I\}$ represents the corresponding observations. Then, the construction of Y_i will depend in the method used. Next, we present three indexing techniques.

In practice, we begin by limiting the total number of variables being tested. Hence, we choose a specific subset of outcomes y_i^m, $m \in \{1, \ldots, M\}$, based on a priori notion of importance. These outcomes will be summarized into one index. This method permits us to avoid the MHT affairs discussed earlier in section 2.3. As mentioned before, the set of outcomes will depend entirely on the experiment context and should be judged on a case-by-case basis.

3.1.1 Anderson (2008)

In his groundbreaking paper, *Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects*, Anderson (2008) proposed an index test. This procedure constructs the index using the inverse of the covariance matrix formed from the transformed variables of interest. Anderson, provides the following steps to implement the corresponding test:

1. For all outcomes, switch signs where necessary so that the positive direction always indicates a "better" outcome.
2. Define J groupings of outcomes (also referred to as areas or domains). Demean all outcomes and convert them to effect sizes by dividing each outcome by its control group standard deviation (denoted by σ_j^{ym} for outcome m in area j). Each outcome y_j^m is assigned to one of these J areas, giving M_j outcomes in each area j, with m indexing outcomes within an area.
3. Create a new variable, AI_{ij}, that is a weighted average of $\tilde{y}_{ij}^m$ for individual i in area j. When constructing AI_{ij}, weight its inputs—outcomes $\tilde{y}_{ij}^m$— by the inverse of the covariance matrix of the transformed outcomes in area j (the outcome weight are called

w_j^m).

$$AI_{ij} = \frac{1}{W_{ij}} \sum_{m \in \mathbb{M}_{ij}} w_j^m \frac{y_{ij}^m - \overline{y}_j^m}{\sigma_{jk}^y} \tag{5}$$

$\hat{\Sigma}_j^{-1}$ is the inverted covariance matrix,

$$\hat{\Sigma}_j^{-1} = \begin{bmatrix} c_{j11} & c_{j12} & \cdots & c_{j1M} \\ c_{j21} & c_{j22} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \ddots \\ c_{jM1} & \vdots & \ddots & c_{jMM} \end{bmatrix} \tag{6}$$

that satisfies $W_{ij} = \sum_{m \in \mathbb{M}_{ij}} w_j^m$. And $\hat{\Sigma}_j$ consists of elements,

$$\hat{\Sigma}_{jkn} = \sum_{i=1}^{N_{jkn}} \frac{y_{ijk} - \overline{y}_{jk}}{\sigma_{jk}^y} \frac{y_{ijn} - \overline{y}_{jn}}{\sigma_{jn}^y} \tag{7}$$

N_{jmn} is the number of observations not missing for both outcome k and outcome n in area j. Now, a simple way to do this is to set the weight on each outcome equal to the sum of its row entries in the inverted covariance matrix for area j. Formally,

$$AI_{ij} = (\mathbf{1}'\hat{\Sigma}_j^{-1}\mathbf{1})^{-1}(\mathbf{1}'\hat{\Sigma}_j^{-1}\tilde{y}_{ij}) \tag{8}$$

Where $\mathbf{1}$ is a column vector of 1's and $\tilde{y}_{ij}$ is a column vector of all outcomes for individual i in area j. Note that this is an efficient generalized least squares (GLS) estimator.[20]

The Anderson index is useful for minimizing the noise resulting from random errors that are uncorrelated across indicators. It also provides an efficient estimation of the treatment effect. Moreover, it offers flexibility to aggregate information from the observed measures that may not be highly correlated or may come from different domains. This process assigns a higher weight to the drawing characteristics that possess more independent information or having lower covariance with respect to other characteristics.

[20] Stata's `swindex` command (Schwab et al., 2021) can be used to obtain the preceding GLS estimator in Stata.

3.1.2 Principal Components (PC)

Principal Component analysis is a statistical method that reduce the dimensionality of data through spectral decomposition. For a more detailed introduction about Principal Component Analysis, refer to Appendix C.

Let C be the $M \times M$ correlation matrix, λ_m and v_m represents the corresponding eigenvalues and eigenvectors for $m \in \{1, \ldots, M\}$:

1. First, get the eigenvectors and eigenvalues from the covariance matrix of the variables of interest. The spectral decomposition of C is given by $C = V\Lambda V' = \sum_{i=1}^{M} \lambda_i v_m v_m'$. The columns of V represents the eigenvectors. They satisfy with orthonormality $v_i' v_j = \delta_{ij}$, and the sign of principal components is not defined. However, the eigenvectors satisfies $\mathbf{1}' v_m > 0$.

2. Next, order the eigenvalues from the smallest to the largest $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_M$. After that, select the eigenvector associated with the larger eigenvalues. These eigenvectors are our Principal Components.

3. The subsequent step is to compute the standardized variables. First, to put us in the right framework, think of PC as a fixed-effects factor analysis with homoskedastic residuals $Z = AL' + E$, where L consists of the larger eigenvectors (also called loadings), the columns of A represents the standardized variables of interest (the component scores), E accounts for the homoskedastic errors and Z is the weighted sum of the component scores and the loadings. Since we are using a correlation matrix, the principal-component scores are in standardized units. So, the standardized variables will be denoted as $\mathbf{a}_m$ with $m = \{1, \ldots, M\}$.

4. Finally, to construct the PC index, select the column of L' that contains the largest eigenvector, v_M, and perform a matrix-vector product with A. Therefore, our index will be,

$$PCI = v_{1M} * \mathbf{a}_1 + v_{2M} * \mathbf{a}_2 + \cdots + v_{MM} * \mathbf{a}_M \tag{9}$$

   Where v_{jM} is the $jth-$element of the $M \times 1$ eigenvector and $\mathbf{a}_m = \left(y_i^m - \overline{y}\right) / \sigma_y^m$ are a $M \times 1$ vector of the standardized variables of interest.[21]

[21] Stata's `pca` command (TX: StataCorp LLC., 2021) can be used to obtain the Principal Components Index in Stata.

3.1.3 *Summary*

The *Summary index* groups all the variables of interest in the simplest way. Denote SI as the Summary index.

1. First, obtain the standardized variables from the outcomes of interest, denoted as $\hat{y}^m = (y_i^m - \overline{y})/\sigma_y^m$.
2. To generate the index, simply add the standardized variables. Suppose there are M outcomes, then the Summary index is calculated as:

$$SI = \sum_{m=1}^{M} \hat{y}^m \tag{10}$$

It is clear that this index assigns the same weight to all variables. The purpose of creating this last index is to use it as a baseline for comparison with the other two more sophisticated procedures. Additionally, some researchers compute the average of the standardized variables instead of simply adding them. For inference purposes, this make no difference since it is just a rescaling by a constant.

3.1.4 Inference

Consider a variable that assigns treatment $Treat_i$ to a fraction of the sample, then, the simplest underlying relationship is

$$Y_i = \tau Treat_i + \varepsilon_i \tag{11}$$

In this framework, often a researcher wants to test the two-sided hypothesis: $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. Then, we can estimate the average treatment effect $\widehat{\tau}$ and its corresponding standard deviation σ. Therefore, the last step for any of the last methods is the same. Regress the the new variable $Y \in \{AI, PCI, SI\}$ on the treatment variable $Treat$. A standard t-test $\widehat{t} = \frac{\widehat{\tau}-0}{s.e(\widehat{\tau})}$, where $s.e(\widehat{\tau}) = \sigma_\tau/\sqrt{I}$ with σ_τ as the sample standard deviation of $\widehat{\tau}$, assesses the significance of the coefficient against zero. Therefore, the statistical problem comes back to that of a simple null hypothesis, and the possible scenarios for hypothesis testing are the same as shown in Table 1.

3.2 Some advantages (and disadvantages) from indexing

The aggregation of related variables into a single index (or a small number of indexes) it is intended to conserve the underlying components of the economic relationship between the outcomes of interest. In principle, reduces the number of hypotheses tested.

Advantages (and possible disadvantages) of indexing rather than MHT procedures are:

1. Ease of general interpretations: Indexes provides a simple and intuitive measure for a set of hypothesis, making it easier for researchers or policymakers to focus on a single, relevant question. In contrast, error correction can be more complex and challenging to understand. On the other hand, magnitudes of index values may also be difficult to interpret, often requiring to consider *effect sizes*.

2. Robustness to over-testing: Since each index represents a single overall test, adding more outcomes to the index does not increas the odds of a false positive (Anderson, 2008).[22] On the contrary, indexing cast a shadow over impacts on specific outcomes of interest and how different groupings can affect the index results.

3. Computational efficiency: Indexes can be more efficient than error correction methods, particularly when dealing with a large set of hypotheses. Error correction methods often require of multiples calculation and adjustment, consuming more time and programming resources. In contrast, creating an index is relatively straightforward, and the statistical evaluation is entirely conventional.

4. Power: Indexes can be more powerful than error correction methods by reducing random error in each outcome measure (Anderson, 2008).

Furthermore, Indexing relies on the ability to classify outcomes as *better* than others, which may not always be clear.

4 Simulations

In this section we construct two main scenarios to evaluate the behaviour of inference on index tests. Much of the literature that uses corrections for MHT and indexing techniques focuses on program evalu-

[22] Several works in the educational field have discussed the issue of simultaneous inference with large numbers of outcomes (Williams et al., 1999).

ation. Therefore, the two main models (presented in this section) mimic situations that require assessing a treatment effect on the index.

4.1 Settings

To assess the performance of the index tests that we are interested, the idea is to construct Monte Carlo experiments in a variety of statistical frameworks. The analyses consider a total of S simulations, I observations and M outcomes in each simulation s. The primary data generating process used is:

$$y_{is}^{m} = \alpha + \tau Treat_{is} + \varepsilon_{is}^{m} \tag{12}$$

Where y_{is}^{m} is the outcome $m \in \{1, \ldots, M\}$ for observation $i \in \{1, \ldots, I\}$ of simulation $s \in \{1, \ldots, S\}$. Moreover, $Treat_{is}$ is a simulated treatment variable, $\varepsilon_{is} \sim N(0, 1)$ is the error component extracted from the Cholesky decomposition of a simulated correlation matrix and ρ is the level of correlation. The use of Cholesky decomposition allows the errors ε_{is}^{m} to be correlated with each other, and therefore, it traspases the correlation to the outcomes. The error distribution is

$$\varepsilon \sim \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \ldots & \rho & 1 \end{pmatrix} \right] \tag{13}$$

The procedure consists of three main steps. First, the outcomes of interest are generated. Second, the indexes are constructed from the outcomes. Finally, assess the statistical significance of the effect of a treatment variable on the index. Moreover, we chose to use ten variables for each index test in all the experiment performed in this thesis. This can be done with any case that satisfies $M > 1$, since as mentioned in Section 3, index test are robust to over-testing.

To start, produce M outcomes following equation (12). To create the variables, first we have to simulate the errors. Note that the Cholesky decomposition provides a lower triangular matrix G, that satisfies $C = GG'$, where C. is the correlation matrix of ε. Then, we create the random variables $b_m \sim N(0, 1)$ with $m \in \{1, \ldots, M\}$. Finally, obtain ε_{is}^{m} as the dot product between the row formed from i-th observation of each b_m and the column vector $\mathbf{g}'_m$, where $\mathbf{g}_m$ is the m-th row from the lower triangular

matrix G. As mentioned before, this procedure generate correlated outcomes.

In addition, we set $\alpha = 1$ and consider two different values of the treatment variable, $\tau = 0$ and $\tau = 0.5$. We also generate another outcome with $\tau = 0.5$ and a random independent error $\epsilon_s \sim N(0,1)$. This additional outcome is created to establish a separate simulated statistical context for multiple comparisons. In summary y_{is}^m is given by:

$$y_{is}^m = \begin{cases} y_{is}^m & if \quad \alpha = 1, \tau = 0, \varepsilon_{is}^m \sim N(0,1), \quad \forall m \in \{1,\dots,10\}, i \in \{1,\dots,I\}, s \in \{1,\dots,S\} \\ yr_{is}^m & if \quad \alpha = 1, \tau = 0.5, \varepsilon_{is}^m \sim N(0,1), \quad \forall m \in \{1,\dots,10\}, i \in \{1,\dots,I\}, s \in \{1,\dots,S\} \\ yr11_{is}^m & if \quad \alpha = 1, \tau = 0.5, \epsilon_s \sim N(0,1), \quad \forall i \in \{1,\dots,I\}, s \in \{1,\dots,S\} \end{cases} \tag{14}$$

Note that each $\varepsilon^m \sim N(0, C)$ with $m \in \{1,\dots,M\}$, are independent from the error used to construct $yr11_{is}^m$, i.e., there are no correlation between $yr11_{is}^m$ and the other outcomes yr_{is}^m.

After generating the outcomes, we construct the three aforementioned index tests for each simulation s, i.e., we simply follow the methods described in Subsection 3.1. We propose four different scenarios to evaluate the performance of each index. The scenarios are the following:

(a) Only null effects (**none**):

- $\alpha = 1, \tau = 0, \varepsilon_i^m \sim N(0,1)$ for every $m \in \{1,\dots,10\}$

(b) Only distinct from zero effects (**all**):

- $\alpha = 1, \tau = 0.5, \varepsilon_i^m \sim N(0,1)$ for every $m \in \{1,\dots,10\}$

(c) A correlated distinct from zero effect (**correlated**):

- $\alpha = 1, \tau = 0, \varepsilon_i^m \sim N(0,1)$ for every $n \in \{1,\dots,9\}$
- $\alpha = 1, \tau = 0.5, \varepsilon_i^m \sim N(0,1)$ for yr_{is}^m, with $m = 10$

(d) An independent distinct from zero effect (**independent**):

- $\alpha = 1, \tau = 0, \varepsilon_i^m \sim N(0,1)$ for every $m \in \{1,\dots,9\}$
- $\alpha = 1, \tau = 0.5, \epsilon_i \sim N(0,1)$ for $yr11_{is}^m$

For simplicity, the scenarios will be called *none*, *all*, *correlated* and *independent*. Now, we will consider every case using y_{is}^m, yr_{is}^m, and $yr11_i^s$. Now, *none* and *all* presents two mainstream cases. The first one was designed to simulate a futile treatment (i.e., does not produced any effect) on the outcomes. In the

second case, there was an actual treatment effect (set at $\tau = 0.5$). For these cases, we expect results consistent with conventional hypothesis testing. Consequently, the null hypothesis for case *none* should be rejected 5% of the time. On the other hand, for case *all* the null should be always rejected.

The last two cases are also of interest. Cases *correlated* and *independent* are an analogous for the FWER presented in Section 2. Since this cases simulates an experiment in which we should at least have one significant hypothesis. The goal is to evaluate the behaviour of the different index tests for a correlated or independent variable that actually received an effect from the treatment.

In this spirit, we could simulate two or more variables that received positive treatment, but selecting just one variable with positive effects offers two advantages. First, as mentioned earlier, FWER represents the probability of making at least one Type I error. Our case thus mirrors the base scenario of this probability and provides a benchmark for comparison upon the usual features of interest. Second, concerning the results between indices, we will be able to assess if differences arise driven solely by the treatment of one outcome. Therefore, unifying our variables of interest and their analysis could be significantly obscured if only one variable receives treatment.

4.2 Empirical performance

In one (of many) MHT-index framework we would like to assess the effect of some treatment (say variable $Treat_i^s$) on an outcome, in this case an index Y_i^s, as we mentioned at the beginning of this section. Hence, to evaluate its performance on inference matters, the natural step is to calculate how many times we get to falsely reject the null hypothesis for each simulation s.

Then, the prevailing and most simple practice is to estimate the following model using OLS estimation:

$$Y_i^s = \beta_0 + \beta_1 Treat_i^s + \varepsilon_i^s \qquad i \in \{1, \ldots, I\} \tag{15}$$

Where Y_i^s is the Anderson, PC or Summary index, $Treat_i^s$ is the treatment variable, and ε_i^s is the error component. The primary objective is to calculate the t statistic associated with the treatment variable. Subsequently, the main goal is to evaluate how many times the null hypothesis H_0 is rejected when it is actually true. The experiment follows the conventional null $H_0 : \beta_1 = 0$, which is rejected if $t > |1.96|$ (so we decide $\alpha = 5\%$ significance level). Furthermore, a t-statistic is obtained for each simulation. Then, all these results are used to create histograms depicting the distribution of the tests. One histogram

for every scenery described in Section 4.

Now, the Monte Carlo experiment described above relies on certain inputs that need to be determined in order to establish appropriate statistical frameworks for MHT. The total number of simulations and observations in each simulation will be set at 1000. As mentioned early, the index will be generated using $M = 10$ variables. Another critical factor in the experiment is the level of correlation ρ, used to generate the variables of interest. Finally, the values of β_0 and β_1 could also be set for different values.

4.2.1 Model 1

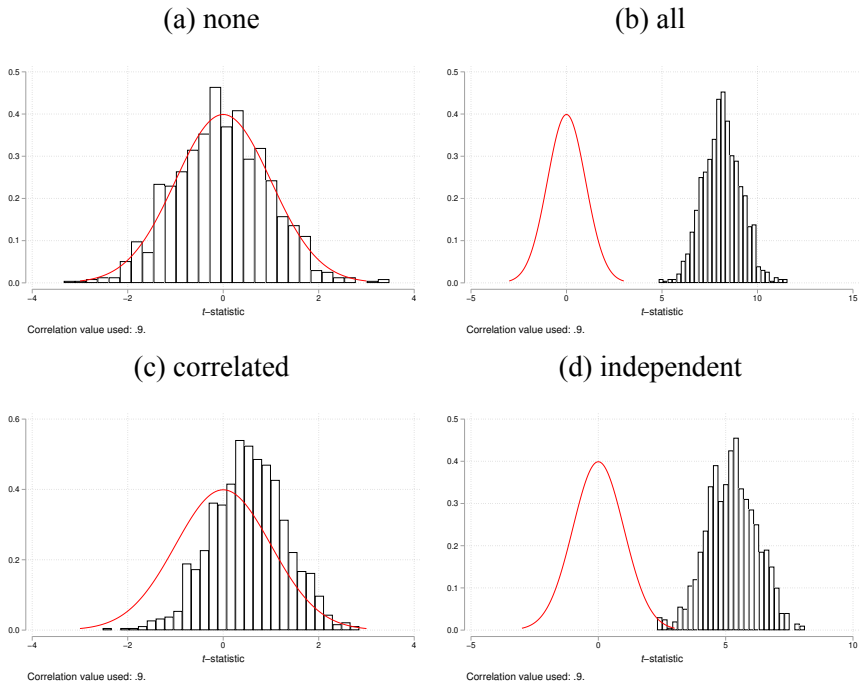
Herein denote as Model 1 the simulations generated following equation 15. This design will perform the simulations for different correlations between variables y_{is} of $\rho \in \{0.01, 0.05, \ldots, 0.95, 0.99\}$.

In first place, Figure 1 show the distribution for the Anderson Index (Figures 2 and 3 present for PC and Summary indexes, respectively), at correlation level $\rho = 0.9$. For case *none*, where the true value of the parameter is $\beta_1 = 0$, the indexes erroneously reject H_0 5% of the time, as we can *observe* in the distribution, which closely resembles a normal distribution. Secondly, case *all*, with a true parameter value of $\beta_1 = 0.5$ the distributions are completely shifted to the right, indicating that the indexes does correctly reject H_0. In the third case *correlated*, the t statistic distribution shift slightly to the right, showing moderate rejection behavior. On the other hand, in scenery *independent*, the indexes have very distinct behaviors. The Anderson index shows a distribution entirely shifted to the right, indicating completely rejection of the null hypothesis. Principal Component Index presents a distribution similar to scenario *none*. In the same way, the Summary index behaves similarly to scenario *correlated*, with moderate rejections. In summary, AI rejects the null in every simulation, PCI rejects a smaller percentage of null hypotheses, and SI rejects a higher proportion compared to weighted methods.

To distinguish more clearly the difference between cases for every value of ρ, results for all the levels of correlation are shown in Table 4, that presents the proportion of rejected tests over the total simulations $S = 1000$. Note that cases *none* and *all* perform in the expected way. For *none* there is a 5% likelihood of rejecting the null, and for *all* there is a 100% of rejecting the null. These results are independent for the level of correlation between variables.

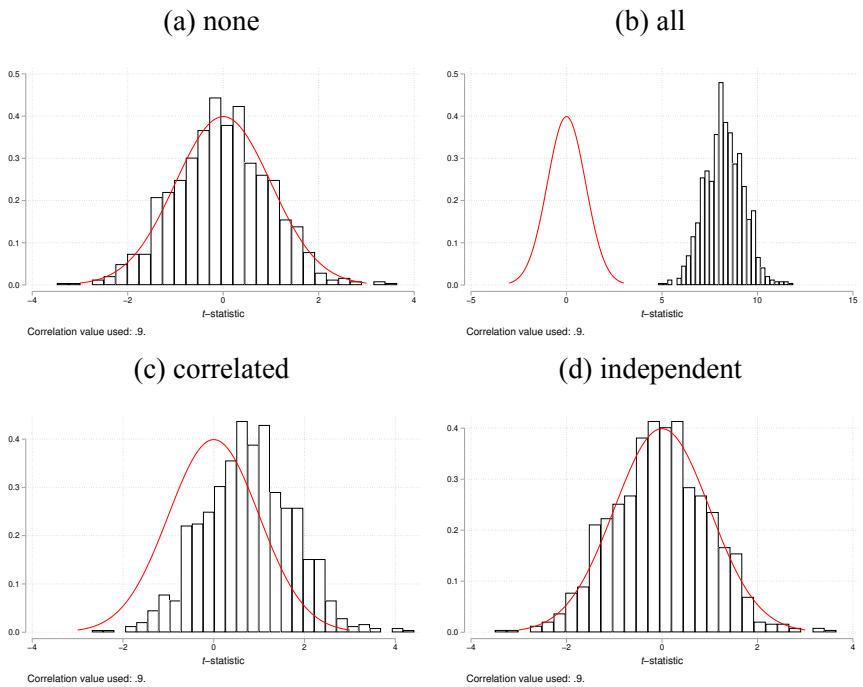
More interesting cases are *correlated* and *independent*. For *correlated*, depending on the level of correlation ρ the proportion of nulls rejected differ widely. Low levels of ρ provokes greater proportions

Figure 1: Performance of Anderson index for $\rho = 0.9$



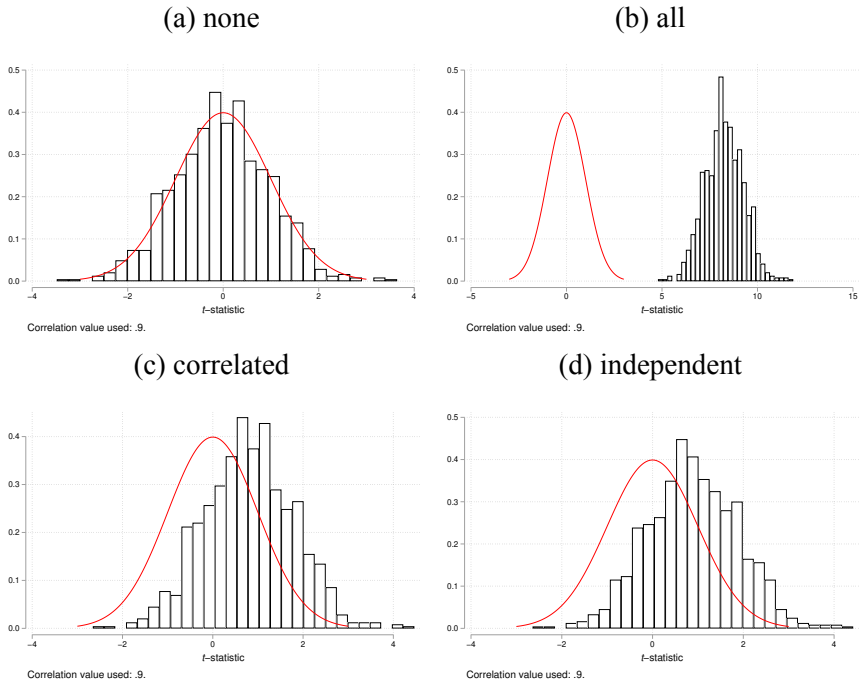
Cases *none* and *all* behaved as expected. In the *correlated* case the t-test distribution slightly shifted to the right. Conversely, the *independent* case distribution is completely shifted away to the right.

Figure 2: Performance of Principal Component index for $\rho = 0.9$



Cases *none* and *all* behaved as expected. In the *correlated* case the t-test distribution slightly shifted to the right. Finally, the *independent* case distribution is pretty similar to the *none* distribution.

Figure 3: Performance of *Summary* Index for $\rho = 0.9$



Cases *none* and *all* behaved as expected. For both the *correlated* and *independent* cases, the t-test distribution slightly shifted to the right.

of rejection. Conversely, high levels of correlation reduces significantly the proportion of tests rejected. AI is the only test that attain a proportion less than 5%, for PCI and SI the proportion is never less than 10%. Finally, case *independent* presents substantial differences depending on the index tests used. AI rejects the null in more than 50% of the tests performed, increasing as the level of correlation goes up. For correlation greater than 60% the proportion of tests rejected tends to 1. Contrarily, PCI performs similar to AI for really low levels of correlation, though the differences arise quickly. However, if the level of correlation grows the proportion of rejected nulls rapidly tends to 5%. Finally, the performance of SI is pretty similar for cases *correlated* and *independent*, following the tendency of less rejection as the correlation increase.

4.2.2 Model 2

An interesting extension from Model 1 is to include a *grouping* variable. In this case, the effect of interest will be the interaction between the treatment variable and the grouping variable. Let $Group_i^s$ be an indicator variable that randomly splits the sample into two groups. The new data generating process for

Table 4: Proportion of tests rejected for the three indexes in each scenario.

	(a)			(b)			(c)			(d)		
ρ	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%	0.047	0.053	0.045	1.000	1.000	1.000	0.623	0.548	0.671	0.675	0.499	0.666
5%	0.056	0.061	0.056	1.000	1.000	1.000	0.496	0.493	0.546	0.813	0.134	0.584
10%	0.058	0.061	0.060	1.000	1.000	1.000	0.382	0.390	0.429	0.866	0.075	0.459
20%	0.042	0.040	0.041	1.000	1.000	1.000	0.281	0.291	0.314	0.940	0.061	0.349
40%	0.053	0.055	0.055	1.000	1.000	1.000	0.156	0.188	0.202	0.988	0.059	0.234
60%	0.049	0.052	0.052	1.000	1.000	1.000	0.119	0.159	0.167	1.000	0.053	0.195
80%	0.055	0.057	0.057	1.000	1.000	1.000	0.091	0.159	0.166	1.000	0.056	0.186
90%	0.050	0.051	0.051	1.000	1.000	1.000	0.030	0.127	0.133	0.999	0.052	0.149
95%	0.038	0.043	0.043	1.000	1.000	1.000	0.009	0.120	0.125	1.000	0.047	0.143
99%	0.052	0.049	0.049	1.000	1.000	1.000	0.000	0.110	0.118	1.000	0.049	0.134

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Case *none* presents only null effects, case *all* only distinct from zero effects, case *correlated* a correlated distinct from zero effect and case *independent* an independent distinct from zero effect. Simulations were done using model (1). The simulated correlation matrix is symmetrical with all off-diagonal values equal.

the outcomes is:

$$y_{is}^{m} = \alpha_1 + \alpha_2 Treat_{is} + \alpha_3 Group_{is} + \tau Treat_{is} \times Group_{is} + \varepsilon_{is} \tag{16}$$

Now we will be interested in the estimation of the interaction between the treatment and the grouping variable, i.e, τ[23]. The experiment performed using the latter equation will be called Model 2. Like Model 1, we execute it for $\rho \in \{0.01, 0.05, 0.1, \ldots, 0.9, 0.95, 0.99\}$. Results are presented in Table 5. For case *none*, at low levels of correlation the proportion of rejection is approximately 5%. But, for greater values of ρ, the null is rejected approximately 6-7% of the time. Therefore, index tests are able to control the odds of committing a Type I error almost as good as in Model 1. For case *all*, at low and medium levels of correlation the results are identical to that of Model 1. Nevertheless, for $\rho > 0.5$, index tests reject the null approximately 95% of the time. For cases *correlated* and *independent*, the performances of PCI and SI are pretty similar to those of Model 1. The difference is that the proportion of rejected nulls are lower and tends to 6-7% as ρ goes higher. At last, AI presents a totally different behaviour in this cases. First, in case *correlated*, the performance is the other way around in comparison to Model 1, i.e., AI rejects more times the null as ρ goes higher. Second, in case *independent*, the proportion of rejected nulls are much lower independently of the level of correlation.

[23] In any case, the researcher could be interested in the parameter α_2. However, in this work we do not analyze that possibility.

Table 5: Proportion of tests rejected for the three indexes in each scenario.

	(a)			(b)			(c)			(d)		
ρ	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%	0.049	0.049	0.048	1.000	1.000	1.000	0.075	0.270	0.227	0.094	0.268	0.226
5%	0.050	0.049	0.049	1.000	1.000	1.000	0.064	0.209	0.184	0.126	0.167	0.183
10%	0.053	0.055	0.056	1.000	1.000	1.000	0.062	0.168	0.159	0.169	0.117	0.153
20%	0.061	0.057	0.058	1.000	1.000	1.000	0.061	0.132	0.124	0.256	0.083	0.128
40%	0.057	0.059	0.058	1.000	1.000	1.000	0.056	0.087	0.090	0.420	0.066	0.104
60%	0.063	0.063	0.063	0.992	0.996	0.996	0.065	0.078	0.080	0.548	0.065	0.093
80%	0.068	0.062	0.062	0.976	0.978	0.978	0.117	0.081	0.082	0.632	0.063	0.090
90%	0.069	0.068	0.068	0.968	0.973	0.973	0.263	0.070	0.075	0.671	0.062	0.079
95%	0.069	0.065	0.065	0.958	0.966	0.966	0.647	0.072	0.071	0.686	0.062	0.074
99%	0.066	0.065	0.065	0.957	0.963	0.963	0.992	0.072	0.074	0.691	0.060	0.071

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Case *none* presents only null effects, case *all* only distinct from zero effects, case *correlated* a correlated distinct from zero effect and case *independent* an independent distinct from zero effect. Simulations were done using model (2). The simulated correlation matrix is symmetrical with all off-diagonal values equal.

The conclusions emerge straightforwardly from the latter results: our indices of primary interest, AI and PCI, exhibit distinct behaviors across all degrees of correlation. This observation also holds true for the indices in cases *none* and *all*. The fairest reasoning suggests that variables should be economically related, but not necessarily statistically correlated. Therefore, the rejection of the null hypothesis could primarily be driven by the level of correlation. This is just one aspect of the story (the choice of index is also crucial), but it suffices to grasp the importance of the analyses conducted here. Next we describe a few more exercises to allow for greater flexibility in the assumptions already set. As will be detailed, the experiment support the same conclusions made here but to a (little) broader extent.

4.2.3 Other extensions

In the models described above we examined the behaviour of the index tests, principally, through variation of the level of correlation. However, the degree of correlation ρ is the same for every off-diagonal value of the correlation matrix C (Equation 13). Nonetheless there are a variety of ways of constructing a correlation matrix. In this subsection, we construct extensions for Model 1.

(a) A first approach is to set different levels of correlation for a sub-matrix of C. Consider the same quantity of simulations $S = 1000$, observations $N = 1000$ and variables $M = 10$. Consequently, these new correlation matrices have 10 rows and 10 columns. Think about two levels of correlation

δ and ρ, now we are interested in the case when δ differ from ρ. In this context, two new inputs that need to be decided. How many variables present a correlation level δ?, and which will be the values of δ and ρ?

To answer the above question, we propose three correlation matrices that allows two, three and four variables to show a correlation level that differs from ρ. The proposed correlation matrices are presented in Appendix B. To evaluate the performance of the index tests in this context we also need to determine the values of ρ and δ. For this extension, we choose a variety of gaps for the correlation levels in each experiment. The combinations are listed in Table 6:

Table 6: Combinations of δ and ρ to simulate different correlations (%)

δ	1	5	10	20	40	60
ρ	99	95	90	80	60	40

Results using the three aforementioned scenarios are presented in Tables A2, A3 and A4, respectively. For the three cases, independently of the combinations used for correlations, cases *none* and *all* have the same expected responses than in Model 1. Hence, the behaviour of the indexes are proper in those scenarios. Results for case *correlated* are a little more tricky. For two and three variables with correlation level δ, AI and SI shows (SI even with four variables) a decay in the proportion of rejection as the gap between δ and ρ decrease. But, if we add a fourth variable, AI behave more erratically increasing the proportion of rejected tests for any combinations of δ and ρ. On the contrary, the PCI presents the contrary behaviour compared to the other two indexes. Ultimately, for case *independent*, AI has constant behaviour, independent of the correlation matrix used, where the level of rejection is at least 80% for any combination of ρ and δ. SI show a considerably lower proportion of rejected tests (never fewer than 20% approximately), that decreases when the divergence gets smaller. But, the most impressive behaviour is performed by the PCI because, independent of the correlation matrix used and also independent of the combinations of correlations, the proportion of rejected tests is always in the range of $4 - 7\%$.

(b) A second extension consider a variation of the SI, and it was presented by Kling et al. (2007). In this case, to create the index, before adding the variables together, all the outcomes have to be transformed into effect sizes (herein known as *Kling Index*, or KI). Therefore, consider M outcomes, all distributed $y_i^m \sim N(\mu_i^m, \sigma_i^m)$ and all observations were classified by a treatment variable T. Two straightforward steps to create this index:

1. Obtain the effect sizes y^{*m}:

$$y_i^{*m} = \frac{y_i^m - \mu_i^{C,m}}{\sigma_i^{C,m}} \tag{17}$$

Where $\mu_i^{C,m}$ and $\sigma_i^{C,m}$ represent the mean and standard deviation of the variable of interest for the control group.

2. Perform a sum of the variables

$$SSI = \sum_{m=1}^{M} y_i^{*m} \tag{18}$$

This indexing techniques will be used to compared with the already presented methods in the replication section.

Ultimately, another parameter set by us is the number of observations N in each sample (for each simulation s). Essentially, this case would be a simply evaluation of a lot of $t-$tests performance for a varying number of observations. This kind of experiment has already been reviewed in simple hypothesis testing literature.[24]

4.3 Statistical Power

Another important matter when performing hypothesis tests is to analyze the power of each test. For this purpose, we carry out a procedure similar to the one exposed in Section 4. Hence, to simulate the power of the index tests, we need to estimate the proportion of correctly rejected null hypotheses. To this end, we define the following variable:

$$propI^s = \begin{cases} 1 & if \quad \textit{p-value} \leq 0.05, \quad \forall s \in \{1, \ldots, S\} \\ 0 & if \quad \textit{p-value} > 0.05, \quad \forall s \in \{1, \ldots, S\} \end{cases} \tag{19}$$

$propI^s$ is the dummy variable for the index $I \in \{AI, PCI, SI\}$ that is equal to one if H_0 is rejected in simulation s (a 5% significance level is used throughout the entire thesis). Moreover, denote the statistical

[24]For the sake of completeness, to evaluate the results of index tests, again consider Model 1 as baseline. Now, we re-examine the Monte Carlo experiment using $S = 1000$ simulations, $M = 10$ variables and vary the number of observations, with values $N \in \{50, 100, 500\}$. Tables A5, A6 and A7 display the results.

power as:

$$PowerI = \frac{\sum_s^S propI^s}{S} \tag{20}$$

$PowerI$ is the proportion of properly rejected null hypotheses for each index test. The numerator is given by the sum of all rejected nulls, and the denominator is the total quantity of simulations.

The procedure also have three main steps: first, generate the variables, construct the index tests and finally assess the statistical significance. Consider Model 1 and let $\alpha = 1$, $\tau \in \{0.01, 0.06, \ldots, 0.91, 0.96\}$ and $\rho \in \{0.01, 0.05, 0.1, 0.2, 0.6, 0.8, 0.9, 0.95, 0.99\}$. Note that we construct the outcomes considering just the case of *effects distinct from zero*. Then, generate y_{is}^m for $m \in \{1, \ldots, M\}$, $i \in \{1, \ldots, I\}$ and $s \in \{1, \ldots, S\}$, with the following parameters $(I, S, M) = (1000, 1000, 10)$. Thus, obtain the three indexes that summarize the information of y_{is}^m, $m \in \{1, \ldots, 10\}$ in the same way as the last subsection. Perform $t-$tests for each simulation and obtain the corresponding p-values. Finally, contrasts the p-values with the standard 5% significance level.

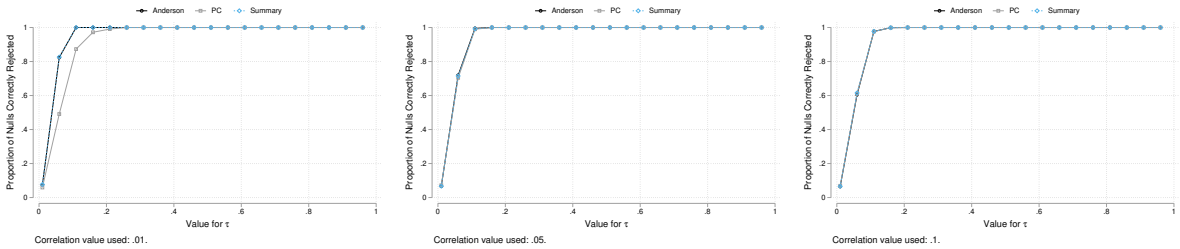
Figure 4 shows the results of the last setting, the proportion of properly rejected hypotheses for the different values of τ and ρ. The proportion of null hypotheses correctly rejected converge rapidly approaches 100%, as the value of τ increases. Nonetheless, the degree of correlation is also of interest to this tendency. With greater value of ρ, increasing statistical power occurs more slowly. But, an important caveat is that, independently of the value of ρ, for $\beta > 0.2$ the power of the index tests is extremely high. At last, the difference in power between the three indexes it seems negligible, except for $\rho = 0.01$ where PCI clearly expose less power.

Note that Figure 4 shows the values of $\beta \in (0, 0.3)$. To present a little more complete display about the power test behaviour for any value of β, Table 7 list the mean of $propI$ at the different levels of correlation ρ, keeping β constant. Although is not strictly accurate measure of power, it shed lights upon the quick changes in power performed by the three indexes.

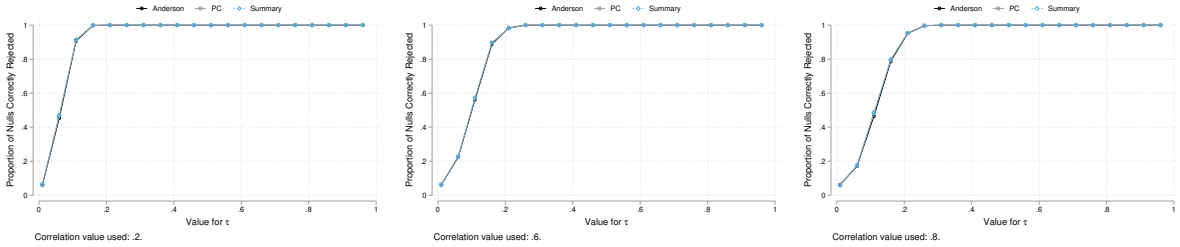
Observe that the right side of Table 7 displays the results for Model 2, which follows the same simulation procedure to obtain the power of the Index tests as described earlier in this Subsection, but considers Equation 16 instead. Also, the plots of power for values of β between 0 and 0.3 are presented in Figure A3. Similarly to the Model 1, greater the value of ρ, the convergence of $\beta \to 1$ is slower. Furthermore, another difference is that the speed of convergence of Model 2 is more moderated, so for this case, the power converge to one for values of $\beta > 0.6$. Finally, notice that AI has slightly less power than PCI and SI, for any correlation level. However, the difference seems negligible.

Figure 4: Power of Index Tests for different levels of correlation using Model (1).

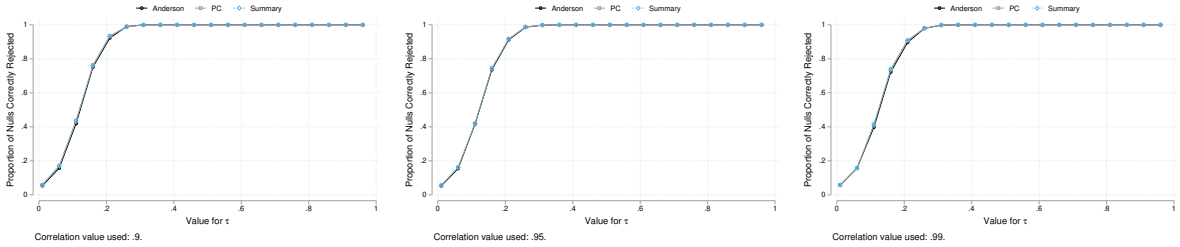
(a) $\rho = 1\%$ (b) $\rho = 5\%$ (c) $\rho = 10\%$



(d) $\rho = 20\%$ (e) $\rho = 60\%$ (f) $\rho = 80\%$



(g) $\rho = 90\%$ (h) $\rho = 95\%$ (i) $\rho = 99\%$



Subfigures (a) to (i) display the power of the three indices for different correlation levels. Greater correlation indicates a slower convergence to maximum power for the three indices. Nevertheless, the maximum power is achieved when the treatment effect is 0.3 or greater for all cases.

Table 7: Mean of $propI$ for any level of correlation ρ, for a given β.

	Model 1			Model 2		
τ	probAI	probPC	probSI	probAI	probPC	probSI
0.01	0.062	0.062	0.063	0.057	0.058	0.058
0.06	0.376	0.347	0.381	0.131	0.144	0.145
0.11	0.684	0.679	0.692	0.318	0.336	0.338
0.16	0.885	0.887	0.890	0.507	0.527	0.528
0.21	0.967	0.969	0.969	0.647	0.671	0.670
0.26	0.995	0.995	0.995	0.770	0.797	0.797
0.31	1.000	1.000	1.000	0.848	0.868	0.868
0.36	1.000	1.000	1.000	0.915	0.931	0.931
0.41	1.000	1.000	1.000	0.957	0.967	0.967
0.46	1.000	1.000	1.000	0.975	0.981	0.981
0.51	1.000	1.000	1.000	0.996	0.998	0.998
0.56	1.000	1.000	1.000	0.997	0.998	0.998
0.61	1.000	1.000	1.000	1.000	1.000	1.000
0.66	1.000	1.000	1.000	1.000	1.000	1.000
0.71	1.000	1.000	1.000	1.000	1.000	1.000
0.76	1.000	1.000	1.000	1.000	1.000	1.000
0.81	1.000	1.000	1.000	1.000	1.000	1.000
0.86	1.000	1.000	1.000	1.000	1.000	1.000
0.91	1.000	1.000	1.000	1.000	1.000	1.000
0.96	1.000	1.000	1.000	1.000	1.000	1.000

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Left side of the table present the results for Model (1), the right side present for Model (2).

For these statistical scenarios, indices converge to maximum power very quickly, but it suggests that we should be mindful of the treatment effect. If the treatment is performing poorly, we may also be losing power. Additionally, power may not be sustained with a similar convergence for various reasons, such as the number of observations or the assumed data generation process. Finally, we can note that the power demonstrated by these indices seems very promising, but it is important to remember that there are many other statistical cases that arise in empirical research.

5 Empirical Applications

To assess the statistical results from the different indexing techniques described throughout this thesis, we replicate two published papers that implemented indices in their respective research. These papers are consistent with our previous analyses using treatment effects on indices. Therefore, each paper obtains treatment effects at some point with a specific indexing technique.

We implement the following steps to evaluate the index performance. First, we replicate the results from the paper related to indexing. Second, we create the other types of indices. Third, we run the exact same regressions as in the main analysis of each paper but using the index left out by the researchers. The main results can be divided in three parts: first, the different magnitudes (and signs), and second, the statistical significance of the treatment effect on the index.

1. **'Acting Wife': Marriage Market Incentives and Labor Market Investment**:

   Bursztyn et al. (2017) studied the relation between marriage market incentives and labor market investments. They conducted a field experiment to assess the behavior of single vs nonsingle women regarding job preferences and skills (their primary experiment). Bursztyn et al. (2017) stated, *"Our main results come from two field experiments that directly test whether single women respond to the studied trade-off by explicitly changing their behavior, making themselves look less professionally appealing."* They utilized the Kling Index as their dependent variable, constructing it by summing 4 variables for this purpose.

   For Table 8 the comparison of interest is across columns. Therefore, each panel represents a different sample set, as implemented in the paper. Two main results are: (1) the sign of the treatment effect is consistent, and (2) the significance level of the treatment is the same, both results do not depend on the indexing techniques used. Nevertheless, a crucial point that Table 8 highlights is

Table 8: Replication Table 4 from (Bursztyn et al., 2017)

	Indexing techniques			
	(1)	(2)	(3)	(4)
Panel A: Single women				
Treatment	-0.56***	-0.79***	-1.15***	-2.94***
	(0.13)	(0.25)	(0.30)	(0.74)
	[0.000]	[0.002]	[0.000]	[0.000]
IndexMean	-0.06	-0.14	-0.22	-0.41
N	59	60	59	59
Panel B: Non-single women				
Treatment	-0.15	-0.19	-0.34	-0.79
	(0.14)	(0.27)	(0.33)	(0.80)
	[0.286]	[0.473]	[0.308]	[0.326]
IndexMean	0.00	-0.00	-0.00	-0.01
N	51	52	51	51
Panel C: Single men				
Treatment	0.04	0.07	0.04	0.21
	(0.12)	(0.21)	(0.27)	(0.66)
	[0.722]	[0.711]	[0.872]	[0.750]
IndexMean	0.15	0.41	0.44	0.99
N	103	104	103	103
Panel D: Non-single men				
Treatment	0.09	0.16	0.16	0.47
	(0.10)	(0.19)	(0.23)	(0.56)
	[0.380]	[0.384]	[0.483]	[0.402]
IndexMean	-0.05	-0.04	-0.07	-0.24
N	130	131	130	130

Standard errors in parentheses, p-values in brackets

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Each column represents a regression performed with a different indexing technique. Model (1) use Kling Index, Model (2) use Anderson Index, Model (3) use Principal Component Index and Model (4) use Summary Index. Each index represent job preferences and skill.

the difference in magnitude depending on the index performed. Clearly, Columns (1) and (2) are similar, which is explained because the Kling Index is the Anderson Index but using equal weights to generate the index. Principal Component Index escapes a little from the first two. Finally, Summary Index shows more extreme magnitudes for every sample.

2. **Iron Deficiency and Schooling Attainment in Peru**:

   In the second paper, Chong et al. (2016) explored the relationship between nutritional deficiencies and intergenerational poverty. Chong et al. (2016) mentioned *"The answer to this question is not only important for evaluating the returns to costly supplementation efforts such as Weekly Iron and Folic Acid Supplementation (WIFS), but is critical for understanding the myriad of ways in which micronutrient deficiencies contribute to poverty and underdevelopment.*" To conduct the experiment they incentivized iron pills consumption among adolescents in rural Peru. For statistical analyses, they constructed the Kling Index as their dependent variable for perceived upward mobility, summing two variables for this purpose. Subsequently, they estimated the following regression:

$$Y_i = a + \beta_1 T_i + wX_i + e \tag{21}$$

   Where Y_i is the outcome of interest, T_i is a binary variable for assignment to the treatment group and X_i is a set of controls.

   For Table 9, the outcomes of interest are across rows. Each column represents a different regression. Here, the results are not as consistent as the above exercise. Depending on the analyzed column, the result for each index can greatly differ. Columns (1) and (3) focus on anemic adolescents. Note that indices appear to be positive, but the Principal Component Index shows a negative impact from the treatment variable. Finally, every index rejects the null at 1%. However, the PC Index is not able to reject at any level. Another interesting feature of Column (3) is that the AI and the SI are not able to reject the null at 1%, similar to the Kling Index. Columns (2) and (4) presents similar results to the ones exposed in the above replication. In this case, the magnitude and sign of each treatment effect go in similar directions. It is worth noting that Columns (2) and (4), according to the literature, are not expected to receive any effect from treatment. Conversely, Column (1) and (3) are the sample of interest since they involve anemic subjects.

As showed above in both replication cases, the different indexing techniques seems to perform appropriately depending on the empirical case, i.e., each test present similar conclusions above treatment

Table 9: Replication Table 6 from Chong et al. (2016)

	Index of perceived upward mobility			
	(1)	(2)	(3)	(4)
Panel A: Kling Index				
Treatment	0.175***	0.008	0.170***	0.025
	(0.062)	(0.058)	(0.064)	(0.058)
	[0.006]	[0.889]	[0.01]	[0.662]
N	81	121	81	121
Panel B: Anderson Index				
Treatment	0.521***	0.059	0.499***	0.110
	(0.198)	(0.187)	(0.206)	(0.188)
	[0.01]	[0.750]	[0.02]	[0.561]
N	81	121	81	121
Panel C: Principal Component Index				
Treatment	-0.269	0.167	-0.311	0.135
	(0.263)	(0.187)	(0.271)	(0.186)
	[0.310]	[0.374]	[0.256]	[0.470]
N	81	121	81	121
Panel D: Summary Index				
Treatment	0.814***	0.087	0.782***	0.166
	(0.306)	(0.289)	(0.318)	(0.291)
	[0.01]	[0.764]	[0.02]	[0.571]
N	81	121	81	121

Standard errors in parentheses, p-values in brackets

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: Each column represents a regression performed with a different set of regressors and different samples. Column (1) regress the treatment variable and a male student identificator on the dependent variable for the anemic part of the sample. Column (2) does the same but for the non-anemic sample. Column (3) add as regressors monthly income, electricity in home and mother's years of schooling for the anemic sample. Column (4) repeat the exercise for the non-anemic sample.

effects on the outcome of interest. However, some differences (that can be subjectively large) arise in terms of magnitude, sign and statistical significance. These contrasts can appear in one or more features.

Concerning to Chong et al. (2016), WIFS programs to improve adolescents health are extremely costly. Then, the statistical significance may be of relevance in public policy. Whether a policymaker wishes to implement the WIFS program or not, since the treatment effects appears to be effective, it would be imperative to note that the statistical significance of the treatment effect lies greatly in the indexing techniques used.

6 Discussion & Conclusion

The analysis conducted in this thesis highlights statistical performance of the different index tests. Our results shows that the choice of the aggregation method is not irrelevant, important features like magnitude or statistical significance can differ greatly. In this final section, we further discuss concerns raised up from our previous results and more.

A remarkable performance is showed by the power of index tests, which, independently of the method used, all present a rapid convergence rate to maximum power, obviously in our statistically-made framework. Therefore, every method proved to be a useful ally to overcome lack of power difficulties that may arise with GERs methods. Nonetheless, other features of index tests raise some concerns, specifically the magnitude and statistical significance of treatment effects.

The issue of magnitude can be tricky, since depending on the method used, the way of interpreting indices and their magnitudes may differ. First, the magnitude will depend on how the dependent variables of interest are transformed, as shown previously, they can be standardized or transform into effect sizes. The literature seems to agree that methods which transform their variables of interest into effect sizes are more appealing to draw conclusions from the magnitude, allowing researchers to have understandable treatment effects. Second, regardless of the transformation upon the dependent variables before generating the index, the resultant treatment effect will normally differ as shown through this thesis. Therefore, presenting results of two or more of the previously discussed indices will not provide a clear understanding of the magnitude of the treatment effect.

Our simulations showed that depending on the context and the index technique used, the statistical significance of the treatment effect can vary considerably. However, there are cases where inference is

consistent through these methods. The more iconic cases visited in this thesis were performed in a FWER fashion, highlighting that, even though this method addresses some concerns regarding the multiplicity of statistical tests by simply reducing the number of tests, is not a definitive answer. As mentioned earlier, the statistical significance issue is not a simple one, and context plays an important role. Therefore, the statistical significance of treatment effects should be carefully take into consideration.

But the final and most important question remains: which indexing techniques should researchers use to fulfill their needs? The answer is rather difficult, but both the Anderson Index and the Principal Components Index seem to perform appropriately depending on the statistical correlation between the variables that form the index. However, this answer is somewhat naive, since in real life there are myriad cases in terms of correlation. Also, as mentioned before, in terms of power, both techniques perform exceptionally well, so this aspect can be set aside. Therefore, a careful approach can be achieved by presenting different indexing methods, as shown throughout this work, to further enhance the corresponding hypothesis.

Now, we can always think deeply about what we are trying to achieve when using indices. Next, we list some practical concerns to help clarify and promote the correct use of this practice.

6.1 Practical concerns

Here, we list some relevant questions to ease the used of index tests in empirical research:

(a) *What kind of conclusion do I want from the index?*

Depending on the subjective relevance that the researcher assigns to their variables of interest, the results from indexing could differ greatly from what is expected. The rejection of the null hypothesis would not necessarily indicate that the treatment had an effect on a large number of the variables of interest; instead, one true effect could be driving the rejection result.

(b) *What variables can I aggregate together?*

A potential drawback is that the index may combine outcomes that are only weakly related and may obscure impacts on specific outcomes that are of interest to particular scholars. Although note that these specific outcomes could also be separately reported for completeness. Questions as: does it make sense to put these variables together? To whom it will be presented the overall effect? can help clarify this issue.

(c) *Thinking carefully about using variables with no variation*

Variation among the variables of interest is essential for generating appropriate weights that constitute the aggregated index. For example, as mentioned by Vyas and Kumaranayake (2006), for PC-based asset indexes we want to avoid asset variables that does not include the necessary variation to include all Socio-Economic Status clusters (e.g. not being able to distinguish between poor and very poor).

Other issues could also be of relevance. For example, what should you do if there are missing values for some components? Exclusion of observations based on missing data from one (or various) variables of interest could significantly lower sample sizes and may lead to bias.

The use of Index Tests has spread quickly in the last decades, with relevant improvements in the techniques in terms of convenience and power. Currently, most authors do not consider carefully what happens when we *unify* our variables of interest and make inference over them, the common practice seems to be performed by habit more than scientific rigor. Moreover, depending on the customs of the field researchers are more inclined to use some techniques than the others. This thesis calls to attention statistical details that should not be ignored, as the results from indexing are deepening policy-related field that can draw erroneous conclusion from this methods.

Finally, there are some questions that fall beyond the scope of this thesis. Here are some important issues that we leave for future research. First, in our simulations, we set only one variable with a positive effect from the treatment, arguing that it resembles FWER's probability. But how might the results change if we add one, two, or three variables with this characteristic? And what is the connection between the number of *treated dependent variables* and the success of using indices?. Second, the empirical applications chosen for this work lack a sufficient number of observations; the sample sizes are rather low. Therefore, further replication could be done with larger datasets to test the performance of the indices and help understand the generality of our results. Lastly, an exhaustive understanding of the mathematical construction of the indices is needed to reveal why the simulations behave so erratically and differently for at least the two main indices of interest, Anderson and Principal Components.

References

K. D. Allee, C. Do, and F. G. Raymundo. Principal Component Analysis and Factor Analysis in Accounting Research. *Journal of Financial Reporting*, 7(2):1–39, 09 2022. ISSN 2380-2154. doi: 10.2308/JFR-2021-005. URL https://doi.org/10.2308/JFR-2021-005.

M. L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, 103(484):1481–1495, 2008.

M. L. Anderson and J. Magruder. Highly powered analysis plans. Working Paper 29843, National Bureau of Economic Research, March 2022. URL http://www.nber.org/papers/w29843.

T. W. Anderson. Asymptotic Theory for Principal Component Analysis. *The Annals of Mathematical Statistics*, 34(1):122–148, 1963.

C. Andrade. Harking, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *J Clin Psychiatry*, (82(1):20f13804), February 2021. doi: 10.4088/JCP.20f13804.

G. Baghumyan. Sexual-Orientation Discrimination and Biological Attributions: Experimental Evidence from Russia. *CERGE-EI Working Paper Series No. 762*, 2023. doi: http://dx.doi.org/10.2139/ssrn.4547312.

L. S. Bender R. Adjusting for multiple testing–when and how? *J Clin Epidemiol*, 54(4):349–9, April 2001. doi: 10.1016/s0895-4356(00)00314-0.

Y. Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721, 2010. doi: https://doi.org/10.1002/bimj.200900299. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.200900299.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

Y. Benjamini, A. Krieger, and D. Yekutieli. Adaptive Linear Step-Up Procedures That Control the False Discovery Rate. *Biometrika*, 93:491–507, 2006.

J. O. Berger and D. A. Berry. Statistical analysis and the illusion of objectivity. *American Scientist*, 76 (2):159–165, 1988. ISSN 00030996. URL http://www.jstor.org/stable/27855070.

D. A. Berry and Y. Hochberg. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1):215–227, 1999. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(99)00044-0. URL https://www.sciencedirect.com/science/article/pii/S0378375899000440.

R. Blakesley, D. Mazumdar, Sati, M. Amanda, P. Houck, G. Tang, r. Reynolds, Charles, and M. Butters. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*, 23(2):255–264, 2009. doi: https://doi.org/10.1037/a0012850.

C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni. Rome*, pages 13–60, 1935.

L. Bursztyn, T. Fujiwara, and A. Pallais. 'acting wife': Marriage market incentives and labor market investments. *American Economic Review*, 107(11):3288–3319, November 2017. doi: 10.1257/aer.20170029. URL https://www.aeaweb.org/articles?id=10.1257/aer.20170029.

S.-Y. Chen, Z. Feng, and X. Yi. A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, 9(6), 2017. ISSN 2077-6624. URL https://jtd.amegroups.org/article/view/13609.

A. Chong and M. Valdivia. Soap Operas and Pro-Savings Attitudes in Poor Rural Areas of Peru. International Center for Public Policy Working Paper Series, at AYSPS, GSU paper2314, International Center for Public Policy, Andrew Young School of Policy Studies, Georgia State University, Aug. 2023. URL https://ideas.repec.org/p/ays/ispwps/paper2314.html.

A. Chong, I. Cohen, E. Field, E. Nakasone, and M. Torero. Iron deficiency and schooling attainment in peru. *American Economic Journal: Applied Economics*, 8(4):222–55, October 2016. doi: 10.1257/app.20140494. URL https://www.aeaweb.org/articles?id=10.1257/app.20140494.

G. Christensen and E. Miguel. Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3):920–80, September 2018. doi: 10.1257/jel.20171350. URL https://www.aeaweb.org/articles?id=10.1257/jel.20171350.

I. Cohen, M. Abubakar, and D. Perlman. Pathways to choice: A bundled intervention against child marriage. *Center for Effective Global Action. University of California, Berkeley.*, pages CEGA Working Paper Series No. WPS–230, 2023. doi: https://doi.org/10.26085/C31C71.

E. K. Denny, D. Dow, G. Levy, and M. Villamizar-Chaparro. Extortion, civic action, and political participation among guatemalan deportees. *British Journal of Political Science*, page 1–20, 2023. doi: 10.1017/S0007123423000418.

B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, Dec. 2001. ISSN 0162-1459. doi: 10.1198/016214501753382129.

D. Evans, S. Cárdenas, and P. Holland. Parent Training and Child Development at Low Cost? Evidence from a Randomized Field Experiment in Mexico. *forthcoming at Journal of Research in Childhood Education.*, 2023.

Food and D. Administration. Multiple endpoints in clinical trials guidance for industry. *Document ID: FDA-2016-D-4460-0024*, 2022. URL https://www.regulations.gov/document/FDA-2016-D-4460-0024.

A. V. Frane. Planned hypothesis tests are not necessarily exempt from multiplicity adjustment. *Journal of Research Practice*, 11:2, 2015. URL https://api.semanticscholar.org/CorpusID:52083818.

J. J. Goeman and A. Solari. Multiple Testing for Exploratory Research. *Statistical Science*, 26(4):584 – 597, 2011. doi: 10.1214/11-STS356. URL https://doi.org/10.1214/11-STS356.

J. J. Goeman and A. Solari. Multiple hypothesis testing in genomics. *Statistics in medicine*, 33(11): 1946–78, 2014. doi: 10.1002/sim.6082.

S. Greenland, S. Senn, R. Kenneth, C. John, P. Charles, G. Steven, and A. Douglas. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, (31(4):337-50), 2016. doi: 10.1007/s10654-016-0149-3.

Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 12 1988. ISSN 0006-3444. doi: 10.1093/biomet/75.4.800. URL https://doi.org/10.1093/biomet/75.4.800.

S. Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

G. Hommel. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386, 06 1988. ISSN 0006-3444. doi: 10.1093/biomet/75.2.383. URL https://doi.org/10.1093/biomet/75.2.383.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

T. Houweling, A. Kunst, and J. Mackenbach. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *International journal for equity in health*, 2(1):1–39, 2003. doi: https://doi.org/10.1186/1475-9276-2-8.

J. R. Kling, J. B. Liebman, and L. F. Katz. Experimental analysis of neighborhood effects. *Econometrica*, 75(1):83–119, 2007.

E. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, Cham, 2022. ISBN 978-3-030-70577-0. doi: 10.1007/978-3-030-70578-7. URL https://link.springer.com/book/10.1007/978-3-030-70578-7.

E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138 – 1154, 2005. doi: 10.1214/009053605000000084. URL https://doi.org/10.1214/009053605000000084.

J. A. List, A. M. Shaikh, and Y. Xu. Multiple hypothesis testing in experimental economics. Working Paper 21875, National Bureau of Economic Research, January 2016. URL http://www.nber.org/papers/w21875.

R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335748.

E. Miguel. Evidence on research transparency in economics. *Journal of Economic Perspectives*, 35(3): 193–214, August 2021. doi: 10.1257/jep.35.3.193. URL https://www.aeaweb.org/articles?id=10.1257/jep.35.3.193.

E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. V. der Laan. Promoting transparency in social science research. *Science*, 343(6166):30–31, 2014. doi: 10.1126/science.1245317. URL https://www.science.org/doi/abs/10.1126/science.1245317.

S. Mullainathan and J. Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, May 2017. doi: 10.1257/jep.31.2.87. URL https://www.aeaweb.org/articles?id=10.1257/jep.31.2.87.

B. A. Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3):61–80, September 2015. doi: 10.1257/jep.29.3.61. URL https://www.aeaweb.org/articles?id=10.1257/jep.29.3.61.

J. P. Romano and M. Wolf. Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73 (4):1237–1282, 2005.

K. J. Rothman. No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1):43–46, 1990. ISSN 10443983. URL http://www.jstor.org/stable/20065622.

B. Schwab, S. Janzen, N. Magnan, and W. M. Thompson. Swindex: Stata module to create a standardized weighted index of multiple indicator variables, 2021. URL https://EconPapers.repec.org/RePEc:boc:bocode:s458912.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions, ii. *Journal of the American Statistical Association,*, 62(626-633), 1967. URL https://books.google.cl/books?id=2_SxMQEACAAJ.

J. Simmons, L. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22 (11):1359–1366, 2011. doi: 10.1177/0956797611417632. URL https://doi.org/10.1177/0956797611417632. PMID: 22006061.

R. Smith, T. Levine, K. Lachlan, and T. Fediuk. The High Cost of Complexity in Experimental Design and Data Analysis: Type I and Type II Error Rates in Multiway ANOVA. *Human Communication Research*, 28(4):515–530, 01 2006. ISSN 0360-3989. doi: 10.1111/j.1468-2958.2002.tb00821.x. URL https://doi.org/10.1111/j.1468-2958.2002.tb00821.x.

T. D. Sterling. Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, 54(285):30–34, 1959. ISSN 01621459. URL http://www.jstor.org/stable/2282137.

J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, (64(3):479–498), 2002.

J. D. Storey. The positive false discovery rate: A bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. ISSN 00905364. URL http://www.jstor.org/stable/3448445.

D. L. Streiner. Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests1. *The American Journal of Clinical Nutrition*, 102(4):721–728, 2015. ISSN 0002-9165. doi: https://doi.org/10.3945/ajcn.115.113548. URL https://www.sciencedirect.com/science/article/pii/S0002916523136963.

D. L. Streiner and G. R. Norman. Correction for multiple testing: Is there a resolution? *Chest*, 140 (1):16–18, 2011. ISSN 0012-3692. doi: https://doi.org/10.1378/chest.11-0523. URL https://www.sciencedirect.com/science/article/pii/S0012369211603401.

A. Tetenov. An economic theory of statistical testing. CeMMAP working papers CWP50/16, Centre for Microdata Methods and Practice, Institute for Fiscal Studies, Sept. 2016. URL https://ideas.repec.org/p/ifs/cemmap/50-16.html.

J. Tukey. The problem of multiple comparisons. 1953.

C. S. TX: StataCorp LLC. Stata multivariate statistics reference manual : Release 17, 2021.

D. Viviano, K. Wuthrich, and P. Niehaus. (when) should you adjust inferences for multiple hypothesis testing?, 2023.

S. Vyas and L. Kumaranayake. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21(6):459–468, 10 2006. ISSN 0268-1080. doi: 10.1093/heapol/czl029. URL https://doi.org/10.1093/heapol/czl029.

R. L. Wasserstein and N. A. Lazar. The ASA's Statement on p -Values: Context, Process, and Purpose. *The American Statistician*, 70(2):129–133, May 2016. doi: 10.1080/00031305.2016.115. URL https://ideas.repec.org/a/taf/amstat/v70y2016i2p129-133.html.

P. H. Westfall and S. S. Young. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. *New York: Wiley*, 1993.

V. S. L. Williams, L. V. Jones, and J. W. Tukey. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69, 1999. ISSN 10769986, 19351054. URL http://www.jstor.org/stable/1165261.

E. Yong. Replication studies: Bad copy. *Nature*, 485(7398):298–300, May 2012. doi: 10.1038/485298a. URL https://ideas.repec.org/a/nat/nature/v485y2012i7398d10.1038_485298a.html.

A Appendix

Figure A1: Figure 1 from Benjamini (2010)

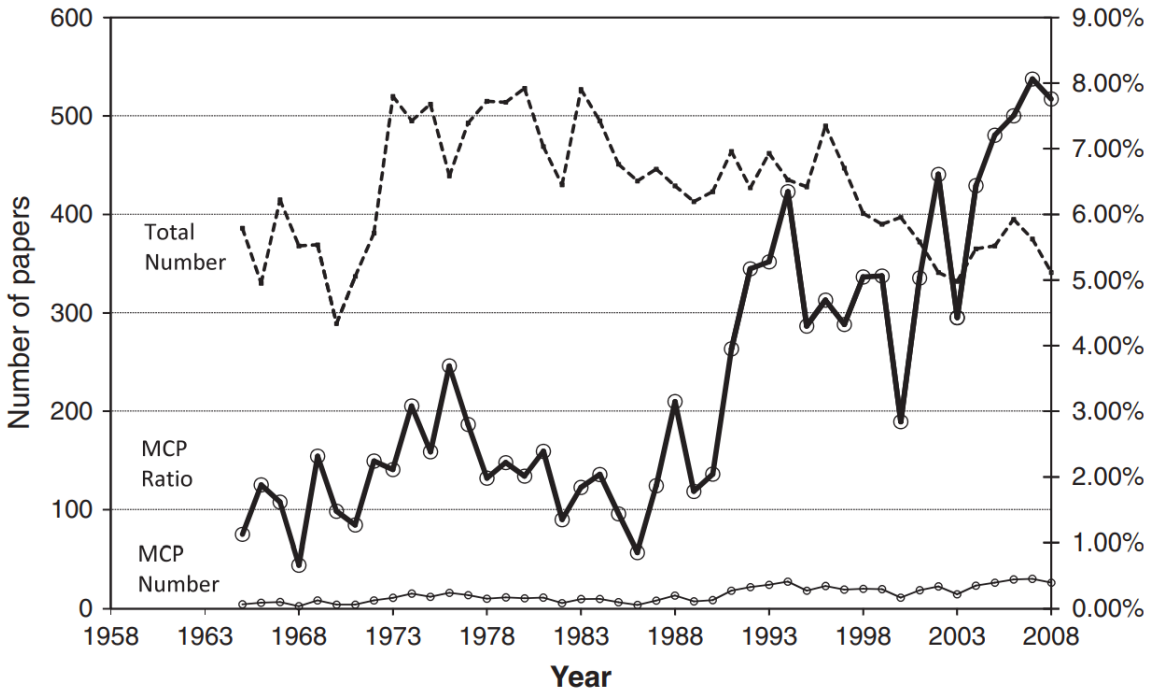
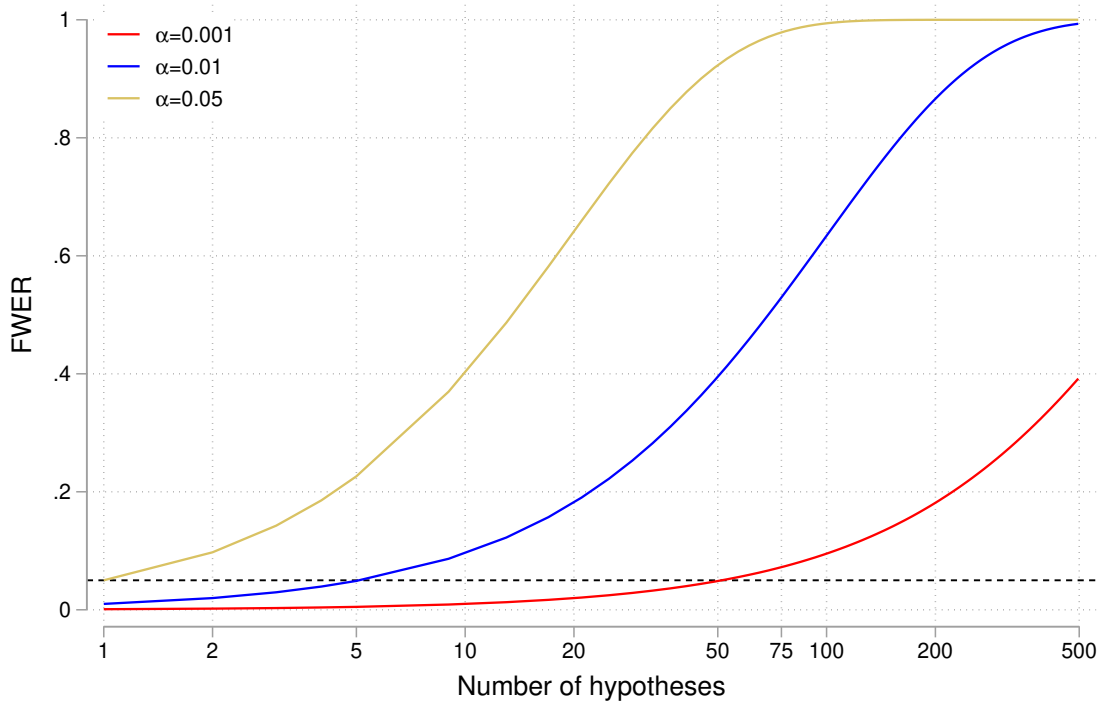


Figure 1 The proportion of papers in the leading four methodological journals that were devoted to simultaneous and selective inference. The total number of papers (left scale) in the *Annals of Statistics*, *Biometrika*, *JASA* and *JRSS B* (dashed line), the number of papers devoted to MCP (slim line) and the proportion (bold lines with empty circles, right scale). Based on computerized search by Johnatan Rosenblat.

B Extensions

$$C_1 = \begin{bmatrix} 1 & \delta_{2,1} & \rho & \dots & \rho \\ \delta_{1,2} & 1 & \rho & \dots & \rho \\ \rho & \rho & \ddots & \vdots & \rho \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho & \dots & \dots & \rho & 1 \end{bmatrix}, C_2 = \begin{bmatrix} 1 & \delta_{2,1} & \delta_{3,1} & \rho & \dots & \rho \\ \delta_{2,1} & 1 & \delta_{2,3} & \rho & \dots & \rho \\ \delta_{3,1} & \delta_{3,2} & 1 & \rho & \dots & \rho \\ \rho & \rho & \rho & \ddots & \dots & \rho \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho & \dots & \dots & \dots & \rho & 1 \end{bmatrix},$$

Figure A2: Likelihood of incurring at least one Type I error



The probability of making at least one Type I error increases with the number of hypotheses tested, and also arises more quickly with a higher level of significance.

Table A1: Multiple testing corrections for FWER and FDR

Methods	Comments
	FWER
Single-step methods	
(Bonferroni, 1935)	1^{st} generation, does not account for correlation among outcomes.
(Šidák, 1967)	Assumes $p-$values mutually independent.
(Tukey, 1953)	Balanced data, ANOVA, $p-$values are not mutually independent.
Stepwise methods	
(Holm, 1979)	Step-down. No dependence assumptions.
(Hochberg, 1988)	Step-up. More power than Holm, 1979.
(Hommel, 1988)	More power than Hochberg, 1998.
(Westfall and Young, 1993)	Step-down, arbitrary dependence, preserve correlation.
(Romano and Wolf, 2005)	Step-down, arbitrary dependence, preserve correlation (Resampling).
	FDR
(Benjamini and Hochberg, 1995)	Provides decision rule (accept/reject) for input values of α. Steup.
(Storey, 2002)	Asymptotic method. More power than BH.
(Benjamini and Yekutieli, 2001)	Provides decision rule for input values of α.
(Benjamini et al., 2006)	More power than Benjamini and Yekutieli, 2001.

Note: Stepdown procedures defines whether the test that looks most significant should be rejected. Stepup procedures looks first at the smallest value of a test statistic when the individual tests reject for large values. Comparisons between FWER techniques can be found in (Blakesley et al., 2009). Explanations and insights about -almost all- error rates in the literature can be found in (Benjamini, 2010).

Table A2: Proportion of tests rejected using C_1.

	(a)			(b)			(c)			(d)		
$\rho(\delta)$	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%(99%)	0.064	0.039	0.059	1.000	1.000	1.000	0.681	0.057	0.613	0.748	0.062	0.620
5%(95%)	0.061	0.043	0.056	1.000	1.000	1.000	0.549	0.097	0.495	0.813	0.062	0.531
10%(90%)	0.062	0.044	0.050	1.000	1.000	1.000	0.460	0.215	0.418	0.899	0.064	0.458
20%(80%)	0.059	0.055	0.059	1.000	1.000	1.000	0.308	0.236	0.294	0.951	0.063	0.330
40%(60%)	0.046	0.054	0.052	1.000	1.000	1.000	0.176	0.188	0.208	0.992	0.066	0.236
60%(40%)	0.055	0.056	0.056	1.000	1.000	1.000	0.079	0.156	0.164	0.999	0.061	0.188

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1) and correlation matrix C_1. The simulated correlation matrix is symmetrical with correlations $\delta_{2,1}, \delta_{1,2}$ different from the rest.

Table A3: Proportion of tests rejected using C_2.

	(a)			(b)			(c)			(d)		
$\rho(\delta)$	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%(99%)	0.053	0.047	0.052	1.000	1.000	1.000	0.748	0.058	0.493	0.801	0.062	0.510
5%(95%)	0.054	0.054	0.052	1.000	1.000	1.000	0.628	0.078	0.435	0.841	0.067	0.441
10%(90%)	0.048	0.059	0.052	1.000	1.000	1.000	0.518	0.099	0.375	0.902	0.068	0.396
20%(80%)	0.049	0.046	0.046	1.000	1.000	1.000	0.367	0.170	0.295	0.961	0.051	0.338
40%(60%)	0.039	0.043	0.042	1.000	1.000	1.000	0.195	0.177	0.200	0.992	0.045	0.223
60%(40%)	0.050	0.053	0.052	1.000	1.000	1.000	0.030	0.169	0.177	0.998	0.053	0.204

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1) and correlation matrix C_2. The simulated correlation matrix is symmetrical with correlations $\delta_{2,1}, \delta_{3,1}, \delta_{1,2}, \delta_{1,3}$ different from the rest.

Table A4: Proportion of tests rejected using C_3.

	(a)			(b)			(c)			(d)		
$\rho(\delta)$	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%(99%)	0.055	0.057	0.061	1.000	1.000	1.000	0.812	0.058	0.392	0.851	0.060	0.397
5%(95%)	0.042	0.056	0.049	1.000	1.000	1.000	0.736	0.064	0.383	0.892	0.061	0.402
10%(90%)	0.042	0.049	0.046	1.000	1.000	1.000	0.565	0.062	0.280	0.920	0.058	0.300
20%(80%)	0.046	0.041	0.046	1.000	1.000	1.000	0.442	0.116	0.272	0.960	0.050	0.303
40%(60%)	0.058	0.057	0.056	1.000	1.000	1.000	0.228	0.159	0.186	0.991	0.055	0.218
60%(40%)	0.051	0.041	0.041	1.000	1.000	1.000	0.999	0.170	0.171	0.995	0.043	0.206

Note: Simulations were performed using $N = 1000$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1) and correlation matrix C_3. The simulated correlation matrix is symmetrical with correlations $\delta_{2,1}, \delta_{1,2}, \delta_{3,1}, \delta_{1,3}, \delta_{3,2}, \delta_{2,3}, \delta_{4,1}, \delta_{1,4}, \delta_{4,2}, \delta_{2,4}, \delta_{4,3}, \delta_{3,4}$ different from the rest.

Table A5: Proportion of tests rejected for the three indexes in each scenario.

	(a)			(b)			(c)			(d)		
ρ	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%	0.046	0.063	0.052	0.991	0.868	0.999	0.071	0.122	0.091	0.068	0.124	0.084
5%	0.052	0.046	0.048	0.968	0.960	0.994	0.061	0.104	0.080	0.070	0.081	0.086
10%	0.048	0.056	0.053	0.937	0.949	0.970	0.056	0.093	0.078	0.096	0.081	0.080
20%	0.051	0.049	0.060	0.822	0.900	0.911	0.056	0.074	0.070	0.108	0.063	0.078
40%	0.033	0.054	0.055	0.561	0.730	0.732	0.029	0.069	0.067	0.127	0.061	0.069
60%	0.032	0.053	0.053	0.401	0.590	0.592	0.025	0.055	0.056	0.126	0.054	0.063
80%	0.025	0.058	0.059	0.285	0.480	0.484	0.012	0.049	0.051	0.163	0.060	0.058
90%	0.022	0.055	0.055	0.217	0.421	0.420	0.008	0.051	0.052	0.144	0.055	0.048
95%	0.015	0.053	0.054	0.214	0.437	0.436	0.002	0.053	0.056	0.151	0.055	0.055
99%	0.018	0.062	0.061	0.197	0.418	0.418	0.000	0.063	0.063	0.171	0.061	0.064

Note: Simulations were performed using $N = 50$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1). The simulated correlation matrix is symmetrical with all off-diagonal values equal.

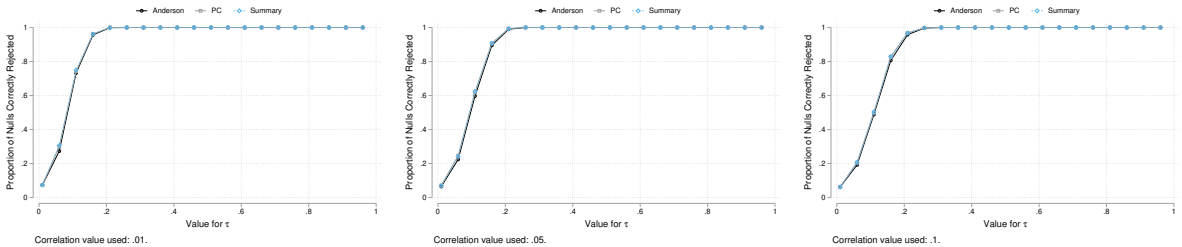
Table A6: Proportion of tests rejected for the three indexes in each scenario.

	(a)			(b)			(c)			(d)		
ρ	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%	0.048	0.048	0.053	1.000	0.983	1.000	0.093	0.161	0.127	0.112	0.168	0.130
5%	0.052	0.051	0.055	1.000	1.000	1.000	0.095	0.130	0.115	0.141	0.112	0.114
10%	0.048	0.048	0.047	0.999	0.999	0.999	0.060	0.088	0.084	0.148	0.075	0.096
20%	0.027	0.041	0.040	0.992	0.996	0.997	0.048	0.068	0.065	0.174	0.051	0.073
40%	0.059	0.071	0.070	0.926	0.958	0.958	0.062	0.086	0.087	0.258	0.073	0.095
60%	0.033	0.058	0.059	0.834	0.878	0.880	0.038	0.071	0.072	0.302	0.066	0.082
80%	0.035	0.060	0.059	0.687	0.772	0.773	0.016	0.062	0.063	0.334	0.065	0.068
90%	0.037	0.067	0.067	0.589	0.697	0.697	0.007	0.064	0.067	0.326	0.070	0.069
95%	0.032	0.052	0.051	0.593	0.715	0.715	0.002	0.059	0.059	0.345	0.052	0.058
99%	0.034	0.050	0.050	0.584	0.689	0.689	0.000	0.061	0.062	0.372	0.052	0.067

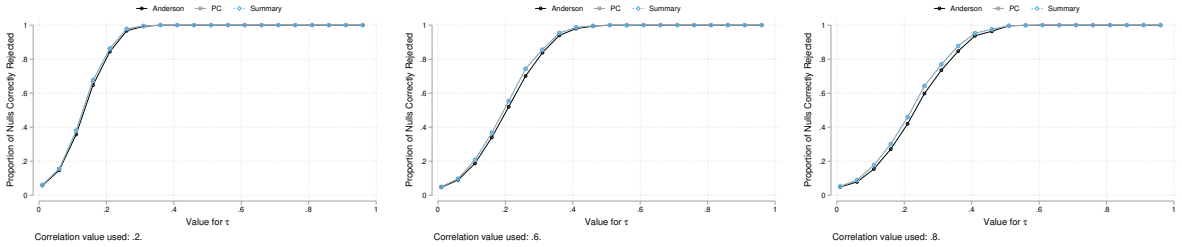
Note: Simulations were performed using $N = 100$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1). The simulated correlation matrix is symmetrical with all off-diagonal values equal.

Figure A3: Power of Index Tests for different levels of correlation using Model (2).

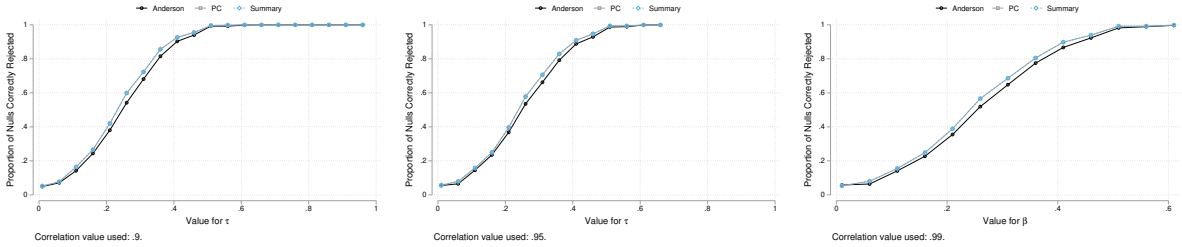
(a) $\rho = 1\%$ (b) $\rho = 5\%$ (c) $\rho = 10\%$



(d) $\rho = 20\%$ (e) $\rho = 60\%$ (f) $\rho = 80\%$



(g) $\rho = 90\%$ (h) $\rho = 95\%$ (i) $\rho = 99\%$



Subfigures (a) to (i) display the power of the three indices for different correlation levels. Greater correlation indicates a slower convergence to maximum power for the three indices. Nevertheless, the maximum power is achieved when the treatment effect is 0.6 or greater for all cases.

Table A7: Proportion of tests rejected for the three indexes in each scenario.

	(a)			(b)			(c)			(d)		
ρ	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI	AI	PCI	SI
1%	0.057	0.052	0.058	1.000	1.000	1.000	0.346	0.402	0.396	0.393	0.398	0.389
5%	0.048	0.051	0.047	1.000	1.000	1.000	0.288	0.326	0.322	0.499	0.131	0.346
10%	0.055	0.052	0.055	1.000	1.000	1.000	0.190	0.228	0.227	0.577	0.084	0.256
20%	0.059	0.054	0.055	1.000	1.000	1.000	0.165	0.177	0.189	0.720	0.072	0.222
40%	0.039	0.043	0.042	1.000	1.000	1.000	0.087	0.120	0.122	0.835	0.049	0.137
60%	0.063	0.067	0.066	1.000	1.000	1.000	0.074	0.124	0.128	0.921	0.067	0.136
80%	0.038	0.046	0.046	1.000	1.000	1.000	0.047	0.104	0.111	0.963	0.048	0.129
90%	0.044	0.046	0.046	1.000	1.000	1.000	0.019	0.072	0.075	0.969	0.047	0.085
95%	0.052	0.054	0.054	1.000	1.000	1.000	0.005	0.077	0.082	0.968	0.055	0.093
99%	0.046	0.052	0.052	0.999	0.999	0.999	0.000	0.076	0.080	0.974	0.054	0.085

Note: Simulations were performed using $N = 500$ observations and $S = 1000$ simulations. Case (a) presents only null effects, case (b) only distinct from zero effects, case (c) a correlated distinct from zero effect and case (d) an independent distinct from zero effect. Simulations were done using model (1). The simulated correlation matrix is symmetrical with all off-diagonal values equal.

$$\mathrm{C}_3 = \begin{bmatrix} 1 & \delta_{2,1} & \delta_{3,1} & \delta_{4,1} & \rho & \dots & \rho \\ \delta_{2,1} & 1 & \delta_{2,3} & \delta_{2,4} & \rho & \dots & \rho \\ \delta_{3,1} & \delta_{3,2} & 1 & \delta_{3,4} & \rho & \dots & \rho \\ \delta_{4,1} & \delta_{4,2} & \delta_{4,3} & 1 & \rho & \dots & \rho \\ \rho & \rho & \rho & \rho & \ddots & \dots & \rho \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \rho & \dots & \dots & \dots & \dots & \rho & 1 \end{bmatrix} \quad (22)$$

C Index Methods

C.1 What is PCA?

In mathematical terms, considering N correlated variables, Principal Components Analysis create uncorrelated components. To better understanding, consider the following system of equations,

$$PC_1 = a_{11}X_1 + \dots a_{1M}X_N \quad (23)$$

$$\vdots \quad (24)$$

$$PC_M = a_{M1}X_1 + \dots + a_{MN}X_N \quad (25)$$

Where a_{MN} represents the weight for the Mth principal component and the Nth variable. The weight for each PC_m ($m \in \{1, \dots, M\}$) are given by the eigenvectors of the correlation matrix C. The variance for each PC is given by the corresponding eigenvalue λ_m. Therefore, ordered eigenvalues prokoves that the largest eigenvector related to the largest eigenvalue explain the most of the variation in the original data. Furthermore, subsequents PCs will be orthogonal and will explain additional but less variation than the first PC.