Quantifying balance for causal inference: An information theoretic perspective

PHD THESIS

Adeola Oyenubi

Student Number: OYNADE001

Thesis presented for the degree of

DOCTOR OF PHILOSOPHY

in the School of Economics

Faculty of Commerce

UNIVERSITY OF CAPE TOWN

Supervised by

Martin Wittenberg and Patrizio Piraino

Acknowledgement

This thesis was made possible because of the support of many people associated with my studies at the University of Cape Town.

I am grateful for the support and guidance provided by my supervisors, Professor Martin Wittenberg and Professor Patrizio Piraino. Their patience in mentoring, encouragement, and putting up with my many errors has contributed immensely to the quality of the study, and my overall PhD experience

I will like to also thank my colleagues for their support and assistance. Specifically, I thank David Fadiran, Jacqueline Mosomi, Love Idahosa, and Chijioke Nwosu for their help in proof reading documents and listening patiently to my arguments when I was developing my proposal.

I am also grateful to my parents and siblings member Engr and Mrs Oyenubi, Adebayo, Abiodun, Adetola, and the Twins (Taiwo and Kehinde Oyenubi) for their help, both in terms of encouragement and financially.

I would also like to thank NIDS for giving me the opportunity to earn a living while battling with my PhD. I appreciate their understanding in allowing me to take time off to attend to my academics, and financial support in terms of funding. I am also grateful to my co-workers at NIDS for their encouragement and support. Specifically I appreciate Timothy Brophy for his Stata advice.

Dedication

To my loving wife, Adetola Oyenubi and my son, Damilare Oyenubi

Contents

Chapter 1. Introduction

1.1 Background and Motivation

Impact evaluation studies are designed to get at the impact of a policy or treatment. The aim is often to assess the viability or success of an intervention. These studies are also useful in cost-benefit analyses for gauging the size of the benefit attributable to an intervention. It is therefore important for an estimated treatment effect to be as free from bias as possible. The object of interest is often the Average Treatment Effect on the Treated (ATT), which focuses explicitly on the effect of treatment on those for whom the intervention is intended. To investigate this, an estimation strategy must solve a missing data problem (Rosenbaum & Rubin, 1983). The central question in this regard is: What would have been the outcome of the treated observations had they not been treated? This is known as the counterfactual outcome. Since the counterfactual outcome is not observable, the estimation process must be able to predict the counterfactual outcome as accurately as possible, from a control sample. In order for a control sample to accurately predict the counterfactual outcome, it must be a good match for the treatment sample. This requires comparability in the covariates of the treatment and the control group, which is referred to as the balancing condition. The ideal way to achieve balance is through a Randomized Control Trial (RCT). This is because, with randomization, the treated and untreated units are drawn from the same population, at random. This ensures that the treatment and control samples have identical distributions of covariates (or are balanced in expectation) in both observed and unobserved covariates. The control sample in this scenario therefore provides the appropriate counterfactual for the treatment group (provided there are no attrition/compliance problems). Randomization is, however, not always possible so that, in most cases, estimation is based on an observational study or quasi-experiment.

The key challenge for observational studies, therefore, is to replicate the kind of result one would expect from a randomized experiment. For this to be successful, a balanced sample is key, because it guarantees that like is compared with like in observables which, by extension, suggests that the same is true for unobservables. This will be the case when the unobservables are correlated with the observables (Imai *et. al,* 2008). It is important to note here that this balance (under randomization) should be in terms of distribution and not just in some moments, like mean and variance. When a control group that balances the

distribution of covariates in the treatment group is used in evaluation, the treatment effect will be unbiased and robust across econometric methods, as one would expect from randomized data. The implication of this is that any inference is based solely on the data, and that it does not rely on model assumption (or model specification). Broadly speaking, imbalance refers to any difference in the distribution of covariates across treatment arms. However, the term imbalance is commonly used to refer to differences in averages (Gelman & Hill, 2007). The evaluation literature in the fields of statistics and economics often uses terms like lack of support/ lack of complete overlap/ violation of common support and imbalance in different ways. For example, Gelman & Hill (2007; chapter 10) refer to two sorts of departures from comparability in the distribution of covariates as imbalance and lack of complete overlap separating the two concepts[1]. They note that imbalance does not necessarily imply lack of complete overlap, and vice versa. Hill & Su (2013) on the other hand, note that failure to satisfy the common support condition can lead to unresolved imbalance (for matching methods), suggesting that one can think of lack of support as a form of imbalance. Furthermore, Imbens & Rubin (2009; chapter 15) refer to lack of support as an extreme case of imbalance. In this thesis, I use the term imbalance to refer to any difference in covariate distributions across treatment arms. This can therefore be differences in mean, variance, differences in other moments apart from the first and second moments, or differences in support. This may manifest as thin/ no support problems (Lechner & Strittmatter, 2009) in finite samples. I note that, if the problem is thin support or no support, it may not become evident when the mean or mean and variance of distributions are compared.

This distinction between balance in distribution and balance in a few moments forms the basis of the questions this thesis attempts to answer. Popular measures of balance compare the first moment (or first and second moments) of covariates across treatment arms. However, a strand of the literature advocates the use of balance measures that consider all parts of the covariate distribution, rather than considering the first moment or the first and second moments only (see Imai *et. al* (2008); Iacus *et. al* (2011); Ho *et. al* (2008); Stuart *et. al* (2010) and Imbens & Rubin (2009) among others). The main argument in support of balance in

[1] The former refers to difference in density on a particular portion of the support while the latter refers to difference in the range of values covered by the support. We say more about this in section 2.1.1.

distribution is that balance is expected to reduce model dependency or sensitivity to model specifications (Imbens & Rubin, 2009; Iacus et. al, 2011, Gelman & Hill, 2007). This is perhaps because such measures are more in line with what randomization achieves under a RCT.

It may be obvious that measures of balance that are based on the mean (and variance) ignore information about balance that may have consequence for bias and robustness of treatment effect estimates. However, it may be less obvious that the kind of distributional measure that is used in assessing balance also matters. This thesis introduces a measure that can be used to quantify imbalance (in both discrete and continuous cases) in covariate distributions across treatment arms. I argue that the proposed measure is more sensitive to imbalance than other measures that are used in the literature. I explore the implication of this balance measure for bias and robustness of treatment effect estimates. The proposed measure should therefore be of interest to researchers doing non-experimental evaluations, because it provides a yardstick for measuring the state or extent of balance compared to a situation where randomized data is available. It may also be of interest in experimental evaluations where one expects balance in distribution. This is because there may be balance concerns after randomization (for example where there is a need to re-randomize) or other concerns like attrition /compliance. Note that randomization ensures that the distribution of covariates is balanced in expectation, so that balance may not be achieved in a given finite sample.

1.2 Core problem and related questions

Estimating the right counterfactual outcome is key in mitigating bias in treatment effect estimation. Let $Y_{0i}\ and\ Y_{1i}$ represent the outcome of each control and treatment unit (i indexes the observations) and $D = 0\ or\ 1$ represent the control and treatment state. The core issue is that the outcome of a unit cannot be observed in both the treated state and the counterfactual state. Therefore the outcome for the control has to proxy for what would have happened to the treated units had they not been treated. In essence, to calculate the ATT, $E(Y_0|D = 0)$ is used as an estimate of $E(Y_0|D = 1)$. This estimate will be correct if the distribution of outcomes in the treatment and control group are similar before treatment. If these distributions are different this may lead to bias (Heckman, Ichimura & Todd, 1997). Since the counterfactual outcome is not observed, covariates of the outcome in the treatment and control groups can be compared, to assess their similarity. The assessment of similarity between the treatment arms is of course an empirical question. Identical (observed and

unobserved) covariate distribution across treatment arms implies that the difference in outcome is attributable to the treatment alone. Furthermore, simple mean difference in outcome will suffice to estimate the casual effect, and this estimate will be robust when other methods are used. In other words, balance creates a situation where controlling for covariates is unnecessary.

If balance is important, then how it is measured is also important. The core research question of this thesis concerns measuring differences in the (observed) distribution of covariates across treatment arms and the implication of how these differences are measured for bias and robustness of effect estimates. I explore whether balance at the mean alone will suffice to recover estimates that are similar to the estimate one would expect from randomized data. I also consider differences in the performance of different distributional measures of balance.

Related to the idea above is the notion of selecting a preferred control group based on its similarity to a given treatment group. This idea is useful in a setting where there is a fixed treatment sample (which means that the ATT is fixed) and a number of plausible control samples that can be used to estimate the treatment effect. For example, in estimating the effect of the National Work Supported (NWS) programme, Lalonde (1986) created 3 control samples by restricting observation by employment history (details in chapter 3). The author therefore had 3 plausible control samples that may have different levels of imbalance (and sample sizes). Under the assumption that better balance (in observables) yields less biased estimates, how can the most appropriate control group be identified in such a setting? Specifically, I am interested in comparing the levels of balance achieved by combining different plausible control samples with a given treatment sample. I argue that a balance measure that is sensitive to all forms of imbalance (as against imbalance in the first two moments should enable me to better assess the extent of similarity between the control samples and a given treatment sample in such a way that the extent of balance (or lack thereof) is more informative about bias and robustness of treatment effect estimates.
Furthermore, I consider the performance of different balance measures under Genetic Matching (GenMatch), which is a method that is used to optimize balance. GenMatch optimizes balance in a sample by weighting the covariates. The process involves improving balance in the matched sample in each successive iteration of GenMatch by changing the

weight attached to the covariates. This process involves improving balance in the matched sample at each stage of the optimization. There is therefore a need to have a balance measure that can adequately distinguish between different levels of balance. I argue that balance measures that focus only on the first or second moments of the distribution of covariates ignore certain aspects of balance, and therefore might not perform well under this method. This thesis therefore compares the performance of a mean based measure with my proposed distributional measure and discusses the difference in results. I note that these differences in results can only be explained by the difference in the measures of balance used, thus highlighting the importance of the measure used to assess balance.

1.3 Balance measures and related issues in the literature

It is well understood in the treatment effect literature that the balancing property is important. However, there are a number of suggestions regarding how to measure balance, and these do not always capture the same idea. As noted earlier, balance often refers to identical first moments in the distribution of covariates in the two treatment arms. This is often accomplished by a t-test of difference in means. However, Imai *et. al* (2008) suggest that rather than limit the comparison to the first moment, one can compare higher order moments of baseline covariates. Ho *et. al* (2008) note that standard deviations of covariates can be compared (in addition to the mean) in assessing balance. What this suggests is that by comparing variance and means one can obtain a broader description of balance, especially for continuous covariates (Austin, 2009). The standardized difference in means is a measure that combines the mean and variance in assessing balance. It compares the difference in means in units of pooled standard deviation. This measure provides a single value (by summarizing information from the first two moments) that can be compared to assess different levels of balance.

There are other proposals in the literature that go beyond the mean and the mean and variance in assessing balance. Austin (2009) suggests comparing quantiles of the covariate distributions (i.e. the minimum, first quartile, median, third quartile and maximum) to allow for broader comparison of distributions of continuous variables. Other proposals that have been put forward include side-by-side box plots (Hoaglin, 1983), empirical cumulative distribution functions (Casella & Berger, 2002; Austin, 2009), quantile-quantile plots (Imai *et. al,* 2008; Ho *et. al,* 2008) and non-parametric density functions (Austin, 2009). What these

measures have in common is that they can provide a broader description of balance, relative to the first two moments. However, their disadvantage is that they only provide a rough idea of balance because they are difficult to compare. That is, they cannot be summarized in a single value like the standardized difference in means. As a consequence, different levels of balance cannot be easily compared with these measures.

There are broad measures of balance that can be easily compared. One such measure that has been used in the evaluation literature is the Kolmogorov-Smirnov (KS) test statistic (see Belitser *et. al* (2011) and Sekhon (2013)). This measure provides a way to compare the distribution of covariates. My proposed measure (the entropic distance metric) also falls into this category, but has not been used for this purpose in the literature. Similar to the KS statistic, it compares distributions, albeit in a different way. In chapter 2, we explore the implication of the difference in the way these measures compare distributions.

One problem with balance measures is that they are often used to assess balance in univariate densities of covariates. However, differences in the joint densities of covariates across treatment arms is the real object of interest in evaluation studies (Iacus *et. al*, 2011). The practise of comparing univariate densities may be informed by the expectation that, if all the univariate densities are balanced, then the joint density will also be balanced. This suggests one disadvantage of mean balance measures. It is more likely that balance in the univariate distribution of covariates will translate into balance in their joint density than it is for balance in means of covariates to yield the same result. One way to summarize information in the joint density is to rely on the results of Rosenbaum & Rubin (1983) and use the propensity score density. In this thesis, balance in propensity score density that satisfies the DW algorithm (Dehejia & Wahba, 1999 & 2002) is used as a proxy for balance in multivariate densities across treatment arms. This provides a uniform baseline for all samples used in this thesis. This baseline allows for the elimination of some imbalance at the edge of the distributions being compared in all samples.

There are two approaches to determining whether a given sample is balanced. Statistical tests of significance, such as t-tests and KS tests, are traditionally used to rule on balance. However, recent studies suggest that it might be better to simply optimize balance (Imai *et. al,* 2008; Ho *et. al,* 2008; Diamond & Sekhon, 2013). Under this approach, there is no stopping rule, balance is improved until it is no longer possible to do so. This is because even a small

imbalance in a variable that is highly correlated with the outcome can lead to a large bias in effect estimate (see Iacus *et al* (2012; section 2.7)). Though the proposed entropic distance can be used as a test statistic, in this thesis it is deployed as a statistic to be optimized to maximise balance. This makes it easier for it to be compared with the standardized difference in means.

1.4 Contribution of the thesis

This thesis introduces the entropic distance metric as a measure that can be used to assess balance in treatment effect estimation. The proposed measure assesses balance as the difference or dissimilarity between (observed) covariate distributions in the groups defined by treatment status. The main argument is that the measure that is used to assess balance is important because of its implication for the inference of treatment effect estimates under various econometric methods. Specifically, ignoring information about certain aspects of balance has consequence for the robustness of treatment effect estimates. This becomes very important for continuous variables, or when the propensity score density is used as a proxy for the joint density of covariates.

The contributions of this thesis can be summarized as follows:

1. A new measure of balance, namely the entropy distance metric is proposed.
2. It is shown that this measure detects balance problems (in observed covariates) that other measures of balance may not detect, and that this difference has consequence for bias and robustness of treatment effect estimates.
3. Related to point 2 above, it is shown that the entropy measure can be used to better identify control samples that result in less biased and more robust treatment effect estimates in a situation where there is more than one plausible control group.
4. Lastly, I argue that this measure can assist in estimating treatment effects more precisely than other measures of balance, when balance is being optimized.

1.5 Thesis Structure

Chapter 2 introduces the entropy measure as a balance measure and discusses its relevance. With a simple discrete example, it is shown that, when comparing two distributions, it is possible to redistribute mass on the densities such that the mean and the KS distance between the distributions remains constant. These differences that are ignored by the mean

and the KS statistic are, however, captured by the proposed entropy metric in a manner that is consistent with what one would want when measuring imbalance. Based on this, I argue that measures that ignore some aspects of balance will not perform well when imbalance manifests as a thin/no support problem in the joint density of covariates.

In chapter 3, I examine the NWS programme, and compare the kind of balance one would expect from a randomized experiment (balance in distribution) and the one often required in observational studies (balance in mean or mean and variance). Based on this, I consider variation in bias and robustness of treatment effect estimates, and the ability of various balance measures to capture this variation. Experimental results are used as the benchmark, and the propensity score density is used as a univariate proxy for the balance in the joint density of covariates. The result shows that distributional measures of balance (entropy measure and the KS statistic) are more correlated with the size of bias, when compared with mean measures of balance (mean and standardized difference in means). Furthermore, the KS statistic is compared with the entropic distance metric. The result suggests that the entropy measure performs better than the KS statistic in quantifying imbalance.

The chapter also considers the relationship between balance and robustness of treatment effect estimates across econometric methods. The result suggests that samples that are balanced in distribution, as measured by the entropic distance, provide more robust treatment effect estimates across econometric methods than samples that are not. By robust we mean lower variability of treatment effect estimates under different econometric methods. My explanation for this result is that imbalance in the sample creates a gap that will be corrected by the econometric approach used in estimating the treatment effect. One can think of various econometric methods as different weighting functions that weight observations to balance the distribution of covariates. Treatment effect estimates across various weighting functions will lead to similar results only when the imbalance in the sample is low (i.e. there is less gap to fill). When this is not the case, treatment effect estimates across methods will be influenced by the weighting approach (or imbalance correction) adopted under the different methods. It is inevitable, therefore, for treatment effect estimates to exhibit variation across methods when there is considerable imbalance in the sample. This thesis shows that samples that are balanced, as measured by the proposed entropy measure,

can be treated like randomized data under the selection on observables assumption. Note that this chapter focuses on balance before matching.

In chapter 4, it is shown that the balance measure used with an algorithm that optimizes balance is important for the result. Using GenMatch, the chapter shows that the size of the effect of the South African Child Support Grant (CSG) on beneficiary's height-for-age z score is lower when the standardized difference in means is optimized, than when the proposed entropy measure is optimized. Given the results shown in the previous chapters, this result suggests that standardized difference in means may be converging at a sub-optimal point in the optimization process. In other words, the treatment effects calculated under the two measures suggest that it may be possible to improve the result of a mean based approach by making use of information in the other parts of the covariate distributions to balance the sample. The balance measure determines the optimal weights under GenMatch, and the weights in turn determine the treatment effect. Using the best balance measure will therefore provide the best estimate of the treatment effect. Across the chapters I also discuss conditions under which I would expect the balance measure not to matter. I, however, argue that these conditions are unlikely to occur in practice. Note that chapter 4 focuses on balance after matching.

Chapter 5 summarizes the results and discusses other possible uses of the entropy measure that are not considered in detail in this thesis. The chapter also discusses some limitations of the proposed measure. One such limitation is that, like other balance measures, the proposed entropy measure cannot tell us anything about balance in unobserved attributes[2]. I do not propose that this measure should necessarily replace existing balance measures. On the contrary, it is a way to augment information provided by other balance measures. It provides a way to distinguish, in a more precise manner, different levels of balance.

[2] Unless the unobserved attribute is correlated with an observed attribute, as noted earlier.

Chapter 2. Quantifying Imbalance in Programme Evaluation

2.1 Introduction

Economists are often interested in the causal effect of a variable on an outcome rather than in the correlation between two variables. The treatment effect literature offers a way to achieve this under some assumptions. Within the treatment effect framework, the population of interest is divided into two groups, units that are exposed to treatment and units that are not. Selecting the control group to be used to estimate the counterfactual outcome of the treatment group is central to the success of any econometric strategy. This can be achieved through a randomized control trial (RCT), or an observational study. Either way, the goal is to achieve balance across the treatment arms so that causal inference is possible. Under an RCT balance is achieved by randomization. Randomization asymptotically, or in expectation, ensures that the distribution of covariates (both observed and unobserved) are identical across treatment arms.

On the other hand, observational studies rely on econometric methods to achieve balance or correct for imbalance. Some methods used in observational studies assume that selection is on observables. Therefore, to estimate unbiased treatment effect under these methods, only observed variables need to be balanced across treatment arms. One way of achieving this goal is by covariate adjustments like weighting, matching or stratifying on propensity scores.

This chapter introduces a new balance measure, namely the entropic distance between distributions. The proposed measure assesses balance in distribution, unlike balance measures that use the first moment (mean only) or the first and second moment (mean and variance). It is argued that the proposed measure captures differences in covariate distribution that other measures may not pick up. To illustrate this point, I provide a simple discrete example where the entropic distance measure detects differences (imbalance) in covariate distribution that other measures of balance used in the literature will not pick up. Since the kind of imbalance in this example has consequences for the bias and robustness of treatment effect estimates, the proposed measure has an important role to play. We compare the proposed measure with the mean and the KS (Kolmogorov-Smirnov) statistic, which is another distributional measure of balance used in the literature (see Belitser *et al* (2011) and Diamond & Sekhon (2013)).

The rest of the chapter is organized as follows. The next section addresses the question of what is meant by imbalance, and departures from comparable distribution of covariates that can be regarded as imbalance. Section 2.2 discusses why balance is important and how it relates to bias. Specifically, the section discusses different components of imbalance. Section 2.3 discusses different measures used to detect imbalance. I argue in this section that distributional measures of balance should be preferred because they are more in line with the balance one would want to obtain from randomized data. Section 2.4 introduces the proposed entropy measure. In section 2.5, a simple example that illustrates the efficacy of the entropy measure is provided. Section 2.6 discusses how the entropy measure can be used, while section 2.7 concludes the chapter.

2.1.1 What is meant by imbalance?

Note that the term imbalance is interpreted in various ways in the literature. The economics literature distinguishes between support problems and imbalance. A support problem refers to a situation where one treatment arm does not have observations (i.e. zero density) on a region of the support while the other treatment arm has positive density on the same region of support. This has to do with differences in the range of values assumed by the covariate(s) across treatment arms. In other words, this problem is a lack of complete overlap of covariate distribution across treatment arms. An imbalance problem, on the other hand, refers to any difference between covariate distributions across treatment arms (Gelman & Hill, 2007). For example, this would apply to a situation where densities are unequal across treatment arms on some portion of the support. In such a case, both treatment arms have positive density on that portion of support. Unequal density could manifest as a thin support problem (in finite samples) where there are few observations on some part of the covariate distribution (Lechner & Strittmatter, 2014). However, some studies refer to a support violation as an extreme case of imbalance. For example, Imbens and Rubin (2015) note that an extreme case of imbalance occurs when the support (range of data values) of covariate distributions across treatment arms differs (also see Gelman & Hill, 2007).

Finding balance can therefore be broken down into two components. The first is ensuring common support (e.g. by dropping observations where empirical density across treatment

arms do not overlap[3]). The second is an additional adjustment to equate densities in portions where the covariate distributions do overlap (Ho *et al,* 2007 sect 6.3).

Balance requires that the joint covariate distributions across treatment arms are identical. However, because there are often many covariates, balance measures in practice compare univariate distributions across treatment arms. Implicitly, this assumes that, when there is balance in each univariate density, this will translate into balance in the joint density. Another method that is used in practice compares quantities that can be used as a proxy for the joint density, e.g. propensity or prognosis scores. For example, when matching on propensity scores, both balance in univariate densities and balance in propensity scores are considered in finding the specification that achieves balance (see for example Dehejia & Wahba (1999 & 2002)). In general, the balancing condition can be written as

$$f(W|D=1) = f(W|D=0) = f(W) \dots\dots\dots (2.1)$$

where $D = \{0\ or\ 1\}$ refers to the treatment status (0 for control group and 1 for treated group), $W = \{w_1, w_2 \dots\dots\dots w_n\}$ is the set of relevant covariates, with n being the number of covariates and f being the density function of W so that $f(W|D=i)$ for $D = 0\ or\ 1$ is the conditional distribution of relevant covariates given the treatment status.

In a case where univariate densities are compared we have

$$f(w_i|D=1) = f(w_i|D=0) = f(w_i) \dots\dots\dots (2.2)$$

where $f(w_i\,|D)$ represents the conditional distribution of the i^{th} covariate given the treatment status. The logic behind comparing univariate densities in assessing balance would assume that if 2.2 is satisfied for all covariates then 2.1 will be satisfied for the same set of covariates.

2.2 Why does balance matter?

Balance diagnostics provide information on the quality of the control group that is used to estimate the treatment effect. This idea is well documented in the matching literature. The

[3] For propensity score matching (PSM), this is achieved by comparing propensity score densities (that achieve balance) and dropping observation that violate the common support condition. King and Zeng (2006) introduce an alternative method to achieve the same objective.

key goal of matching is to prune observations so that the remaining data has better balance in observed covariate distributions across treatment arms (Iacus *et al*, 2012). However, even when matching is not used directly in the estimation process, control observations are often selected (sometimes based on theory) to achieve balance. For example, estimations may be separated by gender or race to make sure that groups that are being compared are comparable. Furthermore, in a regression discontinuity design observations are selected such that they are as close as possible to the point of discontinuity. The point here is that it does not matter if a matching method is used or control samples are selected based on some theory, the main consideration is achieving balance (in both observed and unobserved covariates).

Exactly balanced data means that controlling for differences in covariates is unnecessary because a non-parametric difference in mean outcome will estimate a consistent treatment effect (Iacus *et al*, 2012). Furthermore, treatment effect estimates are often robust across econometric methods and model specification when samples are balanced. A similar argument can be found in Hainmueller (2012)[4]. When observations are not exactly balanced, some model (weighting) will be needed to control for the remaining difference in observed covariates (e.g. parametric regression, weights assigned as a result of propensity score matching or weighting). These models, however, carry certain assumptions. When these assumptions are not true for the data at hand, the result may be biased and/or sensitive to model specification.

2.2.1 Balance and bias

We present our argument by relating equation (2.1) to the characterization given by Heckman, Ichimura, Smith & Todd (1998). The aim is to show in a very general way that balance in the empirical distribution of covariates plays a central role in bias elimination under covariate adjustment techniques in the treatment effect literature (i.e. when we are relying on selection on observables). Under covariate adjustment methods that rely on selection on observables, the average treatment effect on the treated is given by

[4] This author shows that, once the balancing condition is satisfied, estimations become model independent (e.g. in their case, the specification of regression equation becomes irrelevant). Using a simulation, they show that regression results do not vary over a million different specifications when the balancing property is satisfied.

$$ATT = E(Y_1 - Y_0|D = 1) \dots\dots\dots (2.3)$$

$$ATT = E(Y_1|D = 1) - E(Y_0|D = 1) \dots\dots\dots (2.4)$$

(Y_0, Y_1) represent the outcome in the treated and control group. Equation 2.4 shows that this is a missing outcome problem, as we cannot observe the counterfactual mean $E(Y_0|D = 1)$ i.e. the outcome in the control state, given that the unit was treated. Therefore $E(Y_0|D = 1)$ is estimated from the control group i.e. $E(Y_0|D = 0)$. Bias will exist in cases where $E(Y_0|D = 1) \neq E(Y_0|D = 0)$. The bias can therefore be written as

$$Bias = E(Y_0|D = 1) - E(Y_0|D = 0) \dots\dots\dots (2.5)$$

i.e. the difference between the counterfactual mean and the mean of the control group that is supposed to stand in for it. Equation 2.5 can be re-written as[5]

$$Bias = \int_{S_1} E(Y_0|W, D = 1) f(W|D = 1) dW - \int_{S_0} E(Y_0|W, D = 0) f(W|D = 0\,) dW \quad (2.6)$$

where S_1 is the support of W for $D = 1$, S_0 is the support of W for $D = 0$. The bias can then be broken down into components due to differing supports (b), differing distribution of W over the same support in the two populations (c) and differences in outcomes that are present even after controlling for observables (a) (Heckman, Ichimura & Todd, 1997 and Heckman, Ichimura, Smith & Todd, 1998) (see appendix A1 for detailed decomposition).

$$b = \int_{S_1 \backslash S_{10}} E(Y_0|W, D = 1) f(W|D = 1) dW - \int_{S_0 \backslash S_{10}} E(Y_0|W, D = 0) f(W|D = 0) dW$$

$$c = \int_{S_{10}} E(Y_0|W, D = 0)\{f(W|D = 1) - f(W|D = 0)\} dW$$

$$a = \int_{S_{10}} \{E(Y_0|W, D = 1) - E(Y_0|W, D = 0)\} f(W|D = 1) dW$$

where S_{10} is the region over which the support S_1 and S_0 overlaps. A region contained in S_0 but not in S_{10} is denoted $S_0 \backslash S_{10}$ and the region contained in S_1 but not S_{10} is denoted by $S_1 \backslash S_{10}$. According to Heckman, Ichimura & Todd (1997), component (c) is the bias which results from the different distribution of W in the treatment and control groups (difference

[5] Rewriting each mean using the law of iterated expectations, i.e. E(Y|D)=E[E(Y|W D)] where the outermost expectation on the RHS is taken with respect to W.

in densities). Component (b) is bias due to lack of overlap in the covariate distributions, and component (a) is the bias due to violation of selection on observables assumption or Conditional Independence Assumption (CIA).

In terms of the general description of balance that includes non-overlap as a form of extreme imbalance (Imbens and Rubin, 2015), balance (under selection on observables assumption) requires that components b and c of the bias decomposition are equal to zero. To be clear, c refers to imbalance in density as conventionally used while b refers to violation of the common support condition.

When component $b \neq 0$, imbalance (manifesting as lack of overlap) occurs because there are observations in one treatment arm that don't have a counterpart on their region of support in the other treatment arm. This is the same as having observations in the region of support defined by $S_0 \backslash S_{10}$ and $S_1 \backslash S_{10}$. This form of imbalance is often dealt with by restricting estimation to the region of common support. However, in terms of our parameter of interest (ATT) only observations in the region defined by $S_0 \backslash S_{10}$ (control observations) can be dropped in restricting estimation to common support (without redefining the treatment effect); therefore we may still have $b \neq 0$ because of treated observations in the region defined by $S_1 \backslash S_{10}$ that have no counterpart in the control group.

When component $c \neq 0$, imbalance (manifesting as difference in densities) occurs because of shape differences in the region of common support (i.e. S_{10}). Components b and c are linked in that if $f(W|D=1) = f(W|D=0)$ then $S_0 \backslash S_{10}$ and $S_1 \backslash S_{10}$ will be empty. There are, however, different combinations. It is possible to have $c \neq 0$ and $b \neq 0$, especially in observational studies. It is also possible to have $c \neq 0$ but $b = 0$. This can happen in an observational study where component b (of the type $S_0 \backslash S_{10}$) is made zero by restricting estimation to the region of common support and the region defined by $S_1 \backslash S_{10}$ does not exist (either in the middle or the edge of the densities under consideration). Lastly, the only case where components c and b will be zero without having to restrict the estimation to common support is when (2.1) holds in the full sample.

Balance components discussed above have different implications, depending on whether we are considering large or finite samples. In finite samples, non-zero components c and b can

lead to inconsistent results. Therefore, in finite samples thin and no support problems can affect the inference.

2.2.2 Bias and shape difference

Under the assumption that only observations in the region defined by S_{10} need to be considered in estimating treatment effect, we argue in this section that shape differences alone can bias treatment effect estimates[6]. This can occur when estimation is restricted to the region of common support in an observational study. Component c will obviously evaluate to zero if $f(W|D=1) - f(W|D=0) \equiv 0$.

In the case where $f(W|D=1) - f(W|D=0)$ does not evaluate to zero, some bias is likely. It is possible for the bias component c to be zero, for instance if $E(Y_0|W, D=0)$ is constant over common support and the total density in the area of common support is equal. This is because when $E(Y_0|W, D=0) = K$ we can rewrite the expression for c as

$$c = \int_{S_{10}} K\,\{f(W|D=1) - f(W|D=0)\}dW$$

$$= K\left\{\int_{S_{10}} f(W|D=1\,)dW - \int_{S_{10}} f(W|D=0)dW\right\}$$

So if $\int_{S_{10}} f(W|D=1\,)dW = \int_{S_{10}} f(W|D=0)dW$ there would be no bias. However, the assumption of constant outcomes in the absence of treatment is very unrealistic. It is somewhat similar to assuming that treatment effect is constant over the distribution of covariates or propensity scores. Much of the literature has been devoted to discussing the heterogeneity of treatment effects. This is because individuals differ in the way they respond to any kind of treatment. For example, Xie and Brand (2010) show that returns to college education are heterogeneous in that net of observable characteristics individuals who are least likely to obtain college education benefit most from it (see also Smith (2000) and Blundell and Costa (2002)). Therefore, the more realistic assumption is the case where $E(Y_0|W, D=0)$ is heterogeneous over W. Under this condition, the bias due to c will cancel out only if 2.1 is satisfied.

[6] In the next chapter we discuss how shape difference can influence the inference in terms of lack of robustness of effect estimates across methods.

To summarize, imbalance simply means that the distribution of covariates over which the treatment effect is calculated places different weights at different portions of the support. In other words, the factual and the counterfactual are different weighting functions. The implication is that when the values being weighted are a constant over the support, it doesn't matter how you weight. However, when the values being weighted vary, this will not be the case, and bias may result.

2.2.3 Imbalance and matching methods

Matching is an important data processing method that is used in selecting credible counterfactuals in observational studies. Matching methods are valid under strong ignorability (i.e. Conditional Independence Assumption (CIA) or selection on observables and Common Support Assumption (CSA)[7]). Matching is used to pre-process the data in an attempt to mimic experimental conditions. Therefore, the key goal of matching is to prune observations so that the remaining data has better balance in observed covariate distribution across treatment arms. To assess the success of matching methods, balance tests (or checks) are used.

Exact matching of covariates may not be possible because there are too many covariates (curse of dimensionality). Rosenbaum and Rubin (1983) show that, to avoid the curse of dimensionality associated with controlling for many covariates, one can reduce the dimension of a set of covariates $W = \{w_1, w_2 \dots \dots \dots w_t\}$ by conditioning on a balancing score $b(W)$. One such balancing score is the propensity score $P(D = 1|W) = p(W)$, which is the probability of participation given observed covariates. Therefore, instead of matching on W to control for imbalance, matching can be done on $p(W)$.

The CIA requires that treatment assignment D is independent of potential outcomes (Y_0, Y_1) given the propensity scores $p(W) : D \perp (Y_0, Y_1) \mid p(W)$. Another way to put this is to say conditional on $p(W)$, the treatment assignment (D) is as good as randomized (Abadie and Imbens, 2011 pg 2).

The CSA guarantees that each treated unit has a comparable control for matching. It can be written as $0 < p(W) < 1$; i.e. units with the same $p(W)$ values have a positive probability of being observed in both the treatment and control groups.

[7] Both terms are defined below.

Under CIA and CSA, matching is expected to balance the propensity score density i.e. satisfy equation 2.1 (after conditioning on propensity scores). Under these assumptions, a non-parametric estimate of ATT should yield a consistent treatment effect estimate. However, in small samples, this estimate may be sensitive to imbalance (in general) when component b or c of the bias decomposition is not zero. The general argument in section 2.2.1 is related to the PSM method. Under CIA and CSA we expect $D \perp (Y_0, Y_1) \mid p(W)$ or $D \perp W|p(W)$ which forms the basis for balance tests. Here we want the propensity score distributions to be balanced across treatment arms i.e.

$$f(p(W)|D = 1) = f(p(W)|D = 0) \dots \dots . . (2.6)$$

There are two forms of balance tests under PSM. The first is the test that is conducted before matching. This is also referred to as a specification test (Lee, 2013). This test is used to pick the right specification for the propensity score equation. This is what the DW algorithm (Dehejia & Wahba, 1999 & 2002) implements. This algorithm seeks the best way to condition covariates so that both the propensity score density and the covariates are balanced at the mean in blocks defined by the propensity scores. The second test is conducted after matching. It is motivated by concern for balance in the matched sample, and not necessarily in the original or unweighted sample (Lee, 2013). Note that while one may expect 2.6 to be satisfied after matching, it may not be satisfied before matching, even when the DW balance condition is satisfied. Balance diagnostics therefore provide information on how good the propensity score specification is and how successful the matching method is in improving balance in the matched sample. To assess balance in the multivariate distribution of covariates (either before or after matching) the propensity score density can be used based on the result of (Rosenbaum and Rubin, 1983). For example, when dealing with randomized data, one can expect 2.6 to be satisfied before matching.

Alternatively one may be interested in balance in each covariate in the matched sample. For this we require that for each covariate i,

$$f(w_i|D = 1, p(W)) = f(w_i|D = 0, p(W)) \dots \dots . (2.7)$$

in the matched sample.

In the unmatched sample, to deal with component b of the bias decomposition, the sample can be restricted to the region of common support based on the propensity score density. Note that this may not help a lot if $S_1 \backslash S_{10}$ is not empty and we do not want to redefine the treatment effect (this may even occur in the middle of the distribution). As for component c, matching reweights control observations to equate the propensity score density across treatment arms.

2.3 How to detect imbalance: Mean versus Distributional measures

There are numerous tests and checks for balance (Lee, 2013). Examples include the t-test of difference in mean, propensity score specification test (see Dehejia & Wahba (1999,2002)), regression test (Smith & Todd, 2005), test of joint equality of mean, or Hotelling test, and the standardized difference in mean (see Lee (2013) for a review of the tests). These measures compare the first moment of covariates across treatment arms. These balance tests check if

$$E(w_i|D = 1) = E(w_i|D = 0) \dots\dots\dots.. (2.8)$$

Note that this may also be written so that the expression is conditioned on propensity scores (i.e. in the form of 2.7). In contrast to measures that compare distributions, these checks compare the first moment only. There are cases where the second moment is compared in addition to the first moment. As mentioned in chapter 1, this provides a broader description of balance. However, one can go further and compare the distributions so that no information contained in the sample is left out in assessing balance. This is of course a stricter condition. There are cases where mean and distributional measures will be equivalent in terms of their assessment of balance. For example, in the case where the covariates are normally distributed, the distribution can be summarized by the first two moments. Therefore, when covariates are normally distributed, comparing distributions will be equivalent to comparing the mean and variance. In practice, covariate distributions are hardly perfectly normal and sometimes they are discrete, which means that the first two moments may not capture imbalance in a way that is consistent with comparing distributions.

As mentioned earlier, a part of the PSM literature advocates the use of balance measures that consider all parts of the covariate distribution as against considering the first moment or the

first and second moments only (see Diamond & Sekhon (2013); Austin (2009); Stuart *et al* (2013); Sekhon (2007)). Measures that compare distributions have been used for balance assessment in the literature, for example Belitser *et al* (2011) used the Kolmogorov-Smirnov statistic (KS statistic), overlap co-efficient and Levy distance to assess balance. Of these three measures, the KS statistic is the most popular, and is used by other authors such as Diamond & Sekhon (2003) and Stuart *et al* (2013) to assess balance in evaluation studies. The Kolmogorov-Smirnov statistic is defined as the maximum value of the absolute difference between two cumulative distribution functions.

$$KS = \max\{|F(w_0) - F(w_1)|\} \dots\dots\dots\dots\dots (2.9)$$

where $F(w_i)\ i = 0,1$ represents the cumulative distribution function of the covariates being compared. The KS statistic will provide more information about balance relative to mean measures, since it compares distributions. Furthermore, it is safer to assume that 2.1 (balance in joint density) is satisfied when the KS statistic is zero for all covariates than to make the same assumption when mean difference is zero for all covariates.

However, the problem with the KS statistic as a measure of balance is that it has uneven sensitivity to differences in different parts of the distribution (Kaplan & Goldman, 2015). Parizzi & Brcic (2011) note that the KS statistic tends to be more sensitive near the centre of the distribution than the tails (also see Kvam & Vidakovic (2007)). It is reasonable to expect that, given the way the KS statistic is defined, it may be more sensitive to large differences, especially in the centre of the distributions being compared at the expense of smaller difference at the tails. For example, if the distribution contains a point mass, differences in other parts of the distribution may be ignored by the KS statistic. This will have implications for our purpose, since differences in distributions that manifest as thin or no support problems are more likely to occur at the tails. In section 2.5 I provide an example that illustrates this argument, and show that the proposed entropy measure does capture imbalance in cases where the KS statistic fails.

2.3.1 Should a statistical test be conducted to check balance?

Another question is whether it is appropriate to use a formal statistical test as a stopping rule when assessing balance. While there is consensus that balance checks are important, there are two views on how balance should be assessed. Under matching methods, the practice of using hypothesis tests to assess balance has been criticized in the literature. Imai *et al* (2008)

described such formal tests as the "balance test fallacy". This is because such tests are sensitive to sample size. Furthermore, Ho *et al* (2007) argued that balance is a characteristic of the sample and not some hypothetical population. Therefore, hypothesis tests are irrelevant, as interest is on balance in the estimation sample and not in some superpopulation. These authors argue that balance tests do not provide a level below which imbalance can be ignored. This is because if a small imbalance occurs in a variable that has a large effect on the outcome then the small imbalance can translate into a large bias or inefficiency in the treatment effect estimate. These authors (among others) also argue that balance should simply be optimized in the sample at hand (see also Diamond & Sekhon (2013)).

A similar argument is made by Senn (1994, pg 1716) when dealing with randomized experiments. This author noted that the common procedure of conducting balancing tests is "philosophically unsound, of no practical value, and misleading". First, the groups are balanced over all randomizations. Second, for a particular randomization they may be unbalanced. His argument is that the only reason to employ such a test is to examine the process of randomization itself. This is important, because even under randomization there is some chance that there will be imbalance in the sample.

The way chance imbalance is dealt with under randomization suggests a different way of looking at balance. This approach aligns with the one used by authors suggesting that balance should simply be optimized in a sample. One way of dealing with chance imbalance is to re-randomize (Morgan & Rubin, 2012). This involves randomizing and checking for balance until pre-specified balance criteria are meet. However, instead of a formal statistical test (stopping rule approach) the t-statistic itself (or its p-value) is used as a metric to choose the preferred randomization (Bruhn & Mckenzie, 2009). The implementation in chapter 4 is another example of interest in optimizing balance.

In section 2.4.4 I discuss how to compute p-values for the entropy measure. However, in this study I focus on the use of the proposed measure to directly check for balance in the sample, like the way the standardized difference in mean is used to check for balance[8].

[8] This is also similar to the way Belister *et al* (2011) used the distributional measures of balance they examined.

2.4 A new measure of balance in covariate distribution

2.4.1 Imbalance as the entropic distance between covariate distributions

I propose that the extent of imbalance (in actual covariate distributions or in the propensity score density) can be captured by a metric that quantifies the entropic distance between two probability distributions. The requirement that the distance measure should be a metric is important because it allows (under certain conditions) for the assessment of different levels of imbalance. The idea here is that in any application some level of imbalance relative to equation 2.1 (or 2.6) is probably unavoidable (this may be the case even in a randomized experiment). However, insight into how much imbalance there is in a particular application can ensure the results are interpreted properly.

This is a valid approach for estimators that assume that selection is on observables, especially when treatment status defines only two groups. Entropic distance is a metric that is defined over the space of distributions in both continuous and discrete cases (Granger *et al,* 2004). The formal definition of a metric is given as follows (see Schweizer et. al (1960)):

A statistical metric space is an ordered pair $(S, \mathcal{F})$ where S is a set and $\mathcal{F}$ is a metric or a mapping of $d: S\ x\ S \rightarrow \mathbb{R}$ such that for any $p, q, z \in S$ the following holds:

- $d(p,q) \geq 0$ (non-negativity)
- $d(p,q) = 0$ if and only if $p = q$
- $d(p,q) = d(q,p)$ Symmetry
- $d(p,z) \leq d(p,q) + d(q,z)$ Triangle inequality

The discussion below draws from Maasoumi & Racine (2002). Several measures can be used to quantify the distance between distributions. Examples include Shannon's entropy and KL (Kullback–Leibler) distance measures. These measures, however, violate the triangle inequality, and therefore they are not metrics. In my application I use the metric entropy S_ρ proposed by Granger *et al* (2004) as a measure of imbalance. This measure is the normalization of the Bhattacharya-Matusita-Hellinger measure of distance between probability distributions. The measure is given by

$$S_\rho = \frac{1}{2}\int_{-\infty}^{\infty} \left(f_1^{1/2} - f_0^{1/2}\right)^2 \mathrm{dx}$$

for discrete covariates, we have

$$S_\rho = \frac{1}{2}\sum \left(p_1^{1/2} - p_0^{1/2}\right)^2$$

where f_1 and f_0 represent the density of the two distributions being compared (treatment and control, in our case) and p_1 and p_0 represent the mass in the discrete case. The asymptotic distribution of this measure has been derived so it is possible to test if H_0: $S_\rho = 0$ (see Skaug & Tjostheim, 1996). One application of this measure in the economics literature is to measure gender gaps in a way that takes entire earnings distribution into consideration (Maasoumi & Wang, 2012). In this thesis, the entropy measure is implemented with Stata ("Srho" package, Maasoumi & Wang, 2012). It can also be implemented with R using "npdeneqtest" (Li, Maasoumi & Racine, 2009).

The value of S_ρ increases as the dissimilarity (or "distance") between the distributions being compared increases. This measure is a function of the differences in densities and supports of distribution being compared. It will therefore be sensitive to imbalance in terms of both overlap violation and differences in densities on common support. It will also be sensitive to differences in means, variances as well as thin/no support problems.

2.4.2 Brief review on the entropic metric

Measures of distance or dissimilarity between distributions are important because of the role they play in problems of inference and discrimination (Ullah, 1996). The initial concept of distance between distributions was developed by Mahalanobis (1963). Since then, many such measures have been developed (see Rao (1982), Bhattacharyya (1946) and Matusita (1955) among others). For a comprehensive review of distance measures see Cha (2007).

The foundation of the entropic measure being used can be found in information theory. The concept of information refers to a measure of surprise ("surprise" here is a function of the probability density function *pdf* or probability mass function *pmf* of the distribution in question). This surprise is quantified as the logarithm of the inverse of the density at a particular point (see Beck (2009)). Let p be the probability at a given point in a distribution of interest. If the distribution is degenerate all its density will be at the given point. If we were to observe a random draw from this distribution and find its value to be at the point where the entire mass of the distribution is concentrated, our surprise is zero i.e. $(log\frac{1}{p} = log\frac{1}{1} = 0)$. If, however, the value is at a different point, the level of surprise is

infinity ($log\frac{1}{p} = log\frac{1}{0} = \infty$). If we consider a random vector ($y = y_1, y_2, y_3 \dots \dots \dots y_n$) with density function $f(y)$ such that $\int f_y dy = 1$, the measure of information content is given by $\log(f_y)^{-1} = -\log f_y$ (Shannon, 1948).

Entropy, on the other hand, measures the expected value of information (or surprise) over the entire distribution. For a degenerate distribution, entropy is zero, while for a uniform distribution the measure of surprise or information at any point is a non-negative value i.e. ($log\frac{1}{p} > 0$), since this distribution places equal density on every conceivable portion of the support.

Entropy can also be thought of as a measure of uncertainty. Ebrahimi *et al* (1999) compare and contrast variance and entropy as measures of uncertainty. They conclude that, like variance, entropy quantifies the uncertainty of a distribution. However, it quantifies this uncertainty as the deviation of a density from uniform distribution. This idea goes to the core of comparing distributions.

Information theory offers useful concepts to measure the divergence or "distance" between probability distributions. Here we need the concept of relative entropy. Relative entropy between two probability distributions on a random variable is a measure of the distance between them. Formally the relative entropy or the KL distance/divergence between two probability mass functions p and q for random variables $X\ and\ Y$ is given by

$$D(p||q) = \sum_{k=1}^{k} p_k \log(p_k/q_k)$$

$D(p||q)$ reflects the reduction in uncertainty in p as a result of knowledge of q. It is an information theoretic distance of p from q that measures the error in assuming a distribution is q when in fact it is p.

2.4.3 Estimating entropic distance by kernel techniques

Consider the case where there is just one covariate to compare. This can occur when the set W contains only one covariate–i.e. $t = 1$, covariates $W = \{w_1, w_2 \dots \dots \dots w_t\}$ are compared one at a time, or when W is summarized into propensity scores. In practice, implementing the S_ρ measure to compare two distributions involves a two-step procedure. First, the densities to be compared, f_1 and f_0, must be estimated, then the distance between the

estimated densities is measured. Naturally, any error in estimating the densities will filter into the resulting distance measure. Following Granger *et al* (2004) and Maasoumi & Wang (2012) the kernel density estimates of f_1 and f_2 will be used so that

$$\widehat{S_\rho} = \frac{1}{2}\int_{-\infty}^{\infty}\left(\widehat{f_1}^{1/2} - \widehat{f_0}^{1/2}\right)^2 dx$$

where $\widehat{f_1}^{1/2}\ and\ \widehat{f_0}^{1/2}$ are kernel density estimates of f_1 and f_0 respectively. To do this, the choice of bandwidth and kernel becomes important in making sure that the distance measure in the second step is reliable. It turns out that the choice of kernel is not as important as the choice of bandwidth in the kernel density estimation.

The implementation of S_ρ in this study follows the implementation in Maasoumi & Wang (2012). Like these authors, we use the Gaussian kernel and a robust version of the "normal reference rule-of-thumb" bandwidth$\left(= 1.06 \min\left(\sigma, \frac{IQR}{1.349}\right) n^{-\frac{1}{5}}\right)$ where σ is the standard deviation of the variable whose density is being estimated and IQR is the interquartile range. For discrete distributions, we have

$$\widehat{S_\rho} = \frac{1}{2}\sum\left(\widehat{p_1}^{1/2} - \widehat{p_0}^{1/2}\right)^2$$

2.4.4 Computing p-values for the entropic distance

Skaug & Tjostheim (1996) derive the asymptotic distribution of the distance measure (also see Granger *et al* (2004)). However, these asymptotic approximations are known to perform poorly (Maasoumi & Wang, 2012). To test the null of $H_0: S_\rho = 0$ against the alternative, a bootstrap re-sampling method can be used, following Maasoumi & Wang (2012). As highlighted in Maasoumi & Racin (2008), critical values obtained from the asymptotic null distribution do not depend on bandwidth, while the value of the test statistic does. The outcomes of the asymptotic-based tests tend to be sensitive to choice of bandwidth. This will be a problem in applied situations because of competing approaches for data driven bandwidth choice. Therefore, following Granger *et al* (2004), the bootstrap resampling approach can be used.

Consider a sample $W = \{w_1, w_2 \dots \dots \dots w_t; \widetilde{w_1}, \widetilde{w_2} \dots \dots \dots \widetilde{w_t}\}$ of treatment $(w_i)$and control $(\widetilde{w_i})$ observations. The empirical distribution of $\widehat{S_\rho}$ can be constructed under the null of identical

distribution in the treatment and control groups by resampling from the population W. One can then compute percentiles from the ordered bootstrap statistics of $\widehat{S_\rho}$ and use this as a basis for the test for balance.

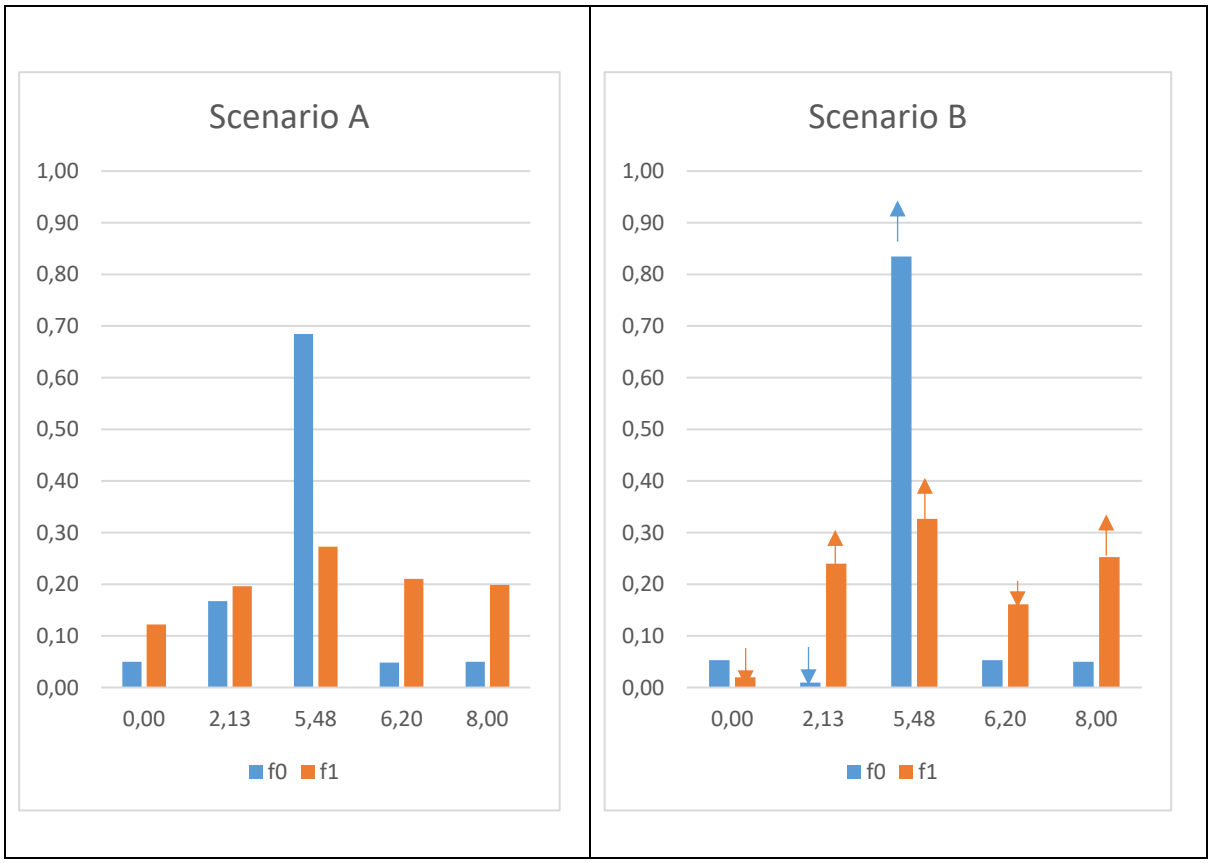
The recommendation is that S_ρ should be used to optimize balance rather than being used as a stopping rule. To do this, one may either optimize the p-value or minimize the entropic distance statistic. In this study I use the latter. This, however, does not address the question of how much imbalance is too much. In a situation where balance is being compared (relative balance) we show that this measure is useful. For cases where we are interested in absolute balance (i.e. how much entropic distance is too much) some care must be taken. This and other caveats are discussed in the concluding chapter (Chapter 5).

2.5 A Discrete example illustrating the efficacy of the proposed entropy

In this section, I present an example that compares the proposed entropic distance with measures that have been used to assess balance in the literature, i.e. KS statistic and the mean. This discrete example shows that there are cases where both the KS statistic and mean balance will miss differences in distributions that the entropic distance will be sensitive to. In this example I redistribute density so that both the KS statistic and the mean remain constant, while the entropy measure picks up the effect of the redistribution. The redistribution has implications in terms of thin and no support problems i.e. components b and c of the bias decomposition (section 2.2.1) in finite samples. The proposed entropy measure should therefore provide more information about balance in such situations compared to the existing measures we examine.

Labels A, B and C represent three scenarios in which two distributions f_0 and f_1 are being compared. The values on the support are (0, 2.13, 5.48, 6.20 and 8). Scenario A is the base case. In scenario B, density is redistributed such that there is higher imbalance than in scenario A. The redistribution is shown by the arrows in figure 2.. In scenario B, a thin support problem is introduced by redistributing mass away from the support region (0.00, 2.13 and 6.20) so that support conditions worsen (at 0.00 and 2.13). The redistributed densities are placed in other regions of support, where there is no thin support problem. This redistribution is such that the mean and the KS statistic remain constant. The KS statistic under scenarios A and B is 0.31 while the difference in mean is zero. The entropic distances are 0.22 and 0.39 in

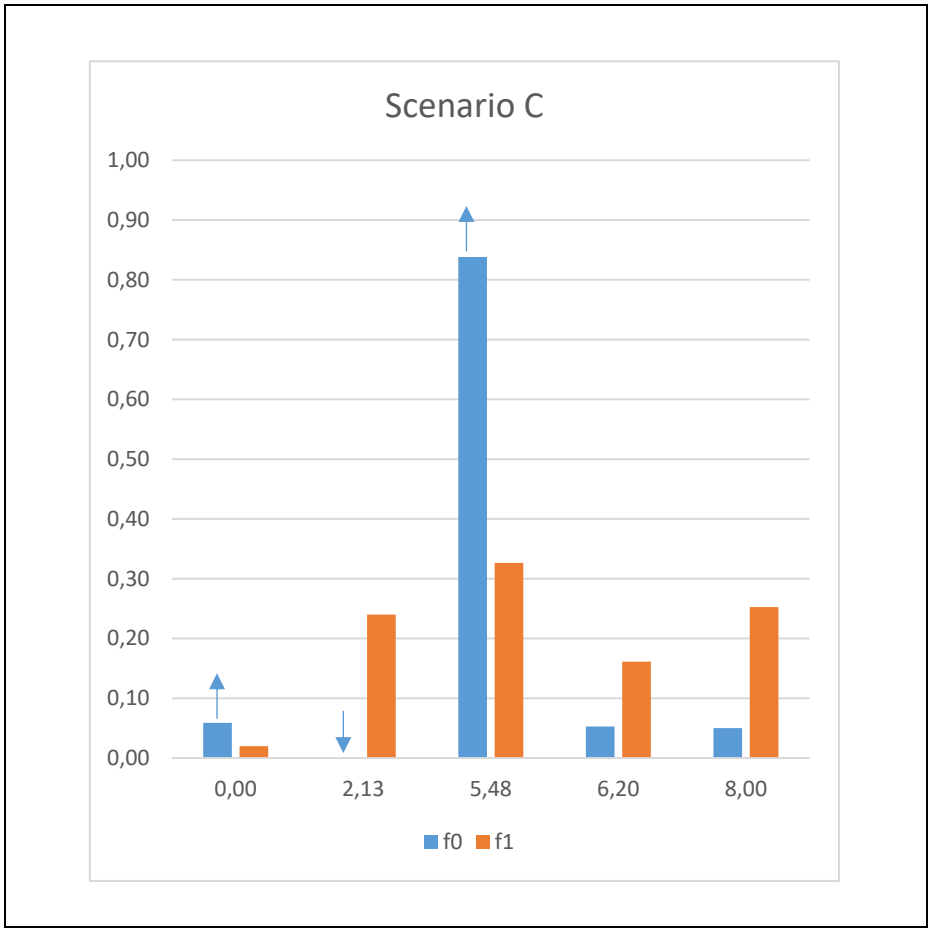
scenarios A and B respectively (see Appendix A2 tables 1-4 for a detailed calculation of the statistics and the tables that generate these results). The balancing condition is better satisfied in scenario A than in B, as suggested by the entropic distance, because of the presence of the thin support problem in scenario B. However, the KS statistic and the means are invariant under this redistribution. The mean and the KS statistic are thus insensitive to imbalance in the form of thin support. Such areas of thin support can bias the treatment effect estimate (Lechner & Strittmatter, 2014).



Finally, in scenario C, density is moved further away from the region of support (2.13) in distribution f_0 such that there is a hole within the interior of the range of values (overlap violation). This was again done so that the KS statistic and mean difference remain constant at 0.31 and 0 respectively.

Under the assumption that less imbalance yields less bias and more robust treatment effect estimates, the relationship between the KS statistic, the mean, and the size of the bias will be distorted in this example. The argument for comparing mean and distributional balance

measures is that distributional balance measures capture aspects of the distribution that mean measures may ignore.



Furthermore, one may assume that when balance in distribution holds for each of the t covariates, balance in joint density of covariates will also hold. It is, however, less plausible that the same logic can be applied to how mean balance translates into balance in joint density. Assuming that balance in mean of each covariate implies balance in joint density across treatment arms, is similar to assuming that other parts of the univariate distribution of covariates do not contribute to the shape and support of the joint density of covariates[9]. Balance in joint density in general requires more than balance in mean or mean and variance.

For example, assume that the covariates can be divided into bins over the support and a multidimensional histogram can be constructed from the set of cells generated by the

[9] Although this argument is valid if all the covariates are normally distributed and they have identical variance, as mentioned earlier.

Cartesian product of the values of the bins. Balance in joint density requires that each cell contains observations from both treatment arms. Furthermore, it requires that each cell contains enough observation from both treatment arms to make within cell estimates credible. When the univariate densities are only balanced at the mean, there is no guarantee that the multivariate density will be balanced.

Given these arguments, distributional measures of balance should be preferred. However, when using distributional measures of balance, care must be taken to use measures that are sensitive to differences in density in all parts of the support. One important difference between the KS statistic and the entropic distance is that the former is based on the maximum distance between the cumulative distribution function, while the latter is based on the sum of all differences between the distribution functions. In terms of assessing balance, this will have implications for the performance of the KS statistic. The key difference is that the KS statistic only responds to the maximum difference between cumulative distribution functions. Consequently, it may ignore differences in other parts of the distribution. As shown in the example, density can be redistributed so that there are thin or no support problems, but the maximum distance between the cumulative distribution remains constant.

Finally, the example also shows the metric property of the entropic distance measure. The measure detects an increasing level of imbalance in a way that is consistent with what imbalance means under treatment effect estimation. This is important when one wants to differentiate between different levels of balance. Using this metric, one can, for example, identify a control sample that better balances the distribution of covariates observed in the treatment group.

2.6 How to use the entropy distance metric

I propose that the entropy distance metric can be used in situations where mean and other measures of balance have been used in the literature. Assessment of balance is often necessary when estimating the treatment effect. To be clear, the proposal is not that mean and variance checks for balance should be discarded. At a minimum, measures that depend on the first two moments should be used (Sekhon, 2007) as they cover basic aspects of balance. My view is that this approach can be augmented by measures that assess balance in higher moments.

The entropy distance can also be used in a re-randomization test as a basis for choosing the preferred randomization. One can, for example, re-randomize until the highest entropic distance between distributions of covariates is lower than some pre-specified value, or the lowest p-value for the entropic distances is above a pre-specified threshold. We expand on this in chapter 5.

The entropy measure can also be used in observational studies. For example: for the PSM method, the entropy distance can be used in the before matching specification test. This is because it is possible to have a situation where more than one specification passes the DW test often used as a heuristic specification check (Lee, 2013). However, the different specifications can lead to propensity score densities (across treatment arms) with different entropic distances between them, and different results in terms of estimated treatment effect. The question will then be how to choose a specification that performs better among the set of specifications that pass the DW test. One way to solve this is to pick the specification that minimizes the entropic distance between the propensity score densities among the set of specifications that pass the DW test. Given our discussion in section 2.5, this specification will be expected to have the lowest imbalance in terms of thin or no support problems.

This argument can be extended to a situation where there is more than one plausible control group for a given treatment group (all of which satisfy the DW test with some appropriate specification). Using the entropic distance between the propensity score density in the unmatched sample, the entropy distance metric can be used to rank such control distributions in terms of their ability to balance the distribution of covariates in the treatment group. I discuss my implementation of this in the next chapter, and show that this may provide valuable information about the size of the bias and the robustness of treatment effect estimates.

The entropy metric can also be used for an after matching test. In this case, it is important to note that the asymptotic distribution of S_ρ will depend on the initial step (creation of the control group through matching). To take this initial step into account, one can bootstrap both steps. It should be noted that the bootstrap often fails for matching estimators. Abadie & Imbens (2008) show this for nearest neighbour matching with replacement. It is therefore not clear if this will work in practice. It may depend on the matching method.

Lastly, there are new matching methods available that avoid the tedious process of continual balance checking and iterative searching over propensity score models. These methods often use some loss function to optimize balance. For example, GenMatch (Diamond & Sekhon, 2013) optimizes balance by reweighting covariates to minimize some loss function (e.g. maximizing the minimum p-value of the t-test and the KS statistic). Instead of using the KS statistic or comparing means, the entropic distance metric can be used. I show how this can be implemented in chapter 4.

2.7 Conclusion

In conclusion, this chapter discusses the implication of imbalance in covariate distributions when estimating the treatment effect. It is noted that for ATT, imbalance can manifest as a thin support problem or a violation of common support in small samples. These may not show up when few moments are compared. The main argument is that the way balance is measured is important. Ideally, for causal inference, one's interest is in comparing the joint distribution of covariates. In the absence of this, distributional measures of balance should be preferred because they provide information about balance in all parts of the univariate distributions being compared. This can more easily translate to balance in joint density than focusing on few moments. Furthermore, distributional measures assess balance in a way that will be consistent with the kind of balance one would expect in a randomized experiment.

This chapter introduces a new measure that can be used to assess balance in distributions and compares it to the KS statistic and the mean, which have been used in the literature to assess balance. I provide an example that shows that the proposed entropic distance is sensitive to imbalance that the mean and KS statistic may be blind to. I discuss various ways in which the proposed measure can be used with existing techniques for data processing like PSM and GenMatch.

Lastly, the proposed measure should not be seen as a competing approach to balance measures that compare mean (or mean and variance). However, in a situation where more information is available, utilizing such information will not harm the goal of assessing balance. On the contrary, it will help to shed more light on the extent of balance in the estimation sample.

3 *Using the Entropic Distance Metric to Rank Control Distributions in Evaluation Studies*

3.1 Introduction

The previous chapter introduced the entropic distance measure as a way of assessing balance. Since balance is central to the success of causal inference, assessing it correctly is important. The entropic measure assesses balance by comparing the distributions themselves. Implicitly, it assesses all the moments of the covariate distributions being compared. This is in contrast to measures of balance that assess only mean and variance. In the chapter, I argue that, in general, distributional measures of balance should be preferred to measures that compare mean and variance only. Furthermore, I argue that the entropic measure captures balance better than the KS statistic (which is another distributional measure of balance) in that it is more sensitive to slight differences in covariate density that can lead to thin or no support problems. In the previous chapter, I demonstrated the efficacy of the entropic measure with a simple example.

In this chapter, I show how the entropy measure can be used to assess balance in real-life data. Ideally, quasi-experiments are supposed to be able to replicate results from randomized experiments. The key to the success of randomized experiments is that they balance the multivariate distribution of covariates across treatment arms. Under the selection on observables assumption, observational studies should replicate experimental results if the control group balances the distribution of covariates in the treatment group. Therefore, in a situation where there are a number of plausible control samples that can be used with a treatment sample, using the control sample that achieves more balance than other plausible control samples will be important in mitigating bias. Since balance measures capture different aspects of balance, in this chapter, I examine how different balance measures perform in identifying control samples that can replicate experimental results. This question is examined under two hypothesis. The first is that increased mismatch or imbalance (between treatment and control samples) will lead to increased bias in treatment effect estimates. Specifically, the results show that the proposed entropy measure predicts bias better than other balance measures. The second is that increased mismatch or imbalance will make estimates more model-dependent (or will make effect estimates vary more across econometric methods).

Specifically, it is shown that the entropy measure predicts variability of effect estimates better than the alternatives. I therefore argue that the proposed entropy measure quantifies balance better than these alternatives.

3.2 Literature review

The link between bias, robustness, and balance is well documented in studies that analyse the National Work Supported (NWS) programme. Lalonde (1986) uses NWS data set to evaluate the performance of non-experimental estimators, using experimental estimates as a benchmark. His results suggest that it is unlikely for an econometrician to recover a treatment effect estimate that is comparable to the one that would have been obtained under a randomized experiment. This is because estimates from observational studies are often biased, and sensitive to the econometric approach and/or model specification. He shows this for simple treatment effects calculated as difference in mean outcome and using standard non-experimental estimators[10] to adjust for selection bias. To do this, Lalonde (1986) selects various theoretically plausible control samples from the Panel Study of Income Dynamics (PSID) and Westat's matched Current population Survey (CPS) datasets. He goes further, to select subsamples of the PSID and CPS data (details on the samples are given in section 3.4). The variability (or lack of robustness) in estimates under different econometric techniques observed by Lalonde (1986) is not necessarily surprising, given that the result from a non-experimental approach depends on the assumptions that validate the approach.

Dehejia & Wahba (1999, 2002) use a subsample (henceforth referred to as the DW Sample) of the data used by Lalonde (1986) (henceforth referred to as the Lalonde Sample). They show that Propensity Score Matching (PSM) can be used to obtain estimates that are more comparable with the experimental estimate under the assumption that selection is on observables. They attribute the success of PSM to its ability to flexibly control for observable differences, by selecting a subset of the PSID and CPS control samples that are more comparable with the treated units in the NWS programme. Their results suggest that adopting the right econometric method can reduce bias.

In contrast, Smith & Todd (2005) find that estimates of the impact of NWS based on PSM are highly sensitive to both the set of variables included in the propensity score equation and the

[10] Lalonde (1986) considers regression, difference-in-difference and latent variable selection approaches.

particular analysis sample used. In their words, PSM does not constitute a "magic bullet" that solves the selection problem in every context. Instead, the goal should be to develop a mapping function from characteristics of data and institutions available in a particular evaluation context to the optimal non-experimental estimator (Smith & Todd, 2005).

Dehejia (2005) points out two key factors that can explain bias and lack of robustness in the PSM results of Smith & Todd (2005). The first is that the propensity score specification should be selected to balance each sample. That is, for every combination of treatment and control observations, a search for a propensity score specification should be conducted. A specification that balances covariates in one sample is not guaranteed to balance covariates in another sample (even when the new sample is a subsample of the initial sample). This can be done with the DW specification test (Dehejia & Wahba 1999, 2002). Second, Ashenfelter (1978) stresses that the failure to match treatment and control groups on pre-treatment labour market characteristics may introduce bias. The author specifically find that observing more than 1 year of pre-treatment earnings is important in estimating the effect of training programs. This is because many people who opt into training programs experience a drop in earnings just prior to opting in. Ashenfelter's refers to this pattern of a drop in earnings prior to program participation, which is commonly observed. This means that if treatment and control group members are not properly matched on pre-treatment labour market characteristics, this may introduce bias.

In summary, the literature suggests that examining balance in the data can help mitigate bias. Our focus is on balance in the data. This is motivated by the fact that when a sample is balanced, the treatment effect is less likely to be biased or vary across econometric methods (Iacus *et al*, 2012). When the data produce a result that is robust across methods, such inference is easier to defend, compared to when data produces a result that lacks robustness across comparable methods. Therefore, under the assumption that selection is on observables, when there are a number of plausible control samples, the control sample that produces robust results should be preferred.

This raises the question of how to find an appropriate (plausible) control sample. This sample should be "more similar" in terms of its covariate distribution to the covariate distribution of observations in the treatment group. This is the gap this study intends to fill. The term "more

similar" is used because some level of imbalance cannot be ruled out in observational studies. One way to deal with some imbalance that may bias treatment effect estimates is to pre-screen[11] the data using the DW algorithm (Dehejia & Wahba, 1999, 2002). This will help reduce the imbalance problem (the kind that manifests as a no support problem) when the sample is restricted to the region of common support based on the propensity score densities in both treatment arms. These propensity scores are generated by the equation that satisfies the balancing condition as defined under the DW algorithm. The DW algorithm can be used to pre-screen data, irrespective of the econometric method one intends to use in estimating the treatment effect. The DW approach deals with the imbalance that manifests as common support violation at the edges of the propensity score density. However, it does not deal with the common support violation that may appear in the middle of the propensity score density. More importantly, this approach (restricting estimation to common support, as implemented in for example, Becker & Ichino *et al* (2002)) cannot be used to deal with the imbalance that manifests as a thin support problem. That is, in cases where there are few observations in one treatment arm compared to the other treatment arm, either in the middle or at the edges of the propensity score density. Regions of thin support can lead to finite sample bias and increased variance (Lechner & Strittmatter, 2014). In general, areas of thin or no support may increase bias and variance of estimators (Kahn & Tamer, 2010, Crump et al, 2009).

Lechner & Strittmatter (2014) analyse several practical adjustment procedures that have been proposed in the literature to deal with the problem of thin support or no support in the context of PSM. These procedures rely on rules to drop observations so that the population for which the treatment effect is estimated is the one for which the distribution of covariates overlaps adequately. The problem with this approach is that the treatment observation may be dropped so that the treatment effect is redefined, and the new treatment group will depend on the available control units in the control sample being used. While this is a valid approach, this thesis considers an alternative that may be valuable where there is more than one plausible control sample that can be used with a given treatment sample.

The implication of the above is that, even when a given combination of treatment and control samples satisfies the DW algorithm, there may still be cause for concern in terms of different

[11] By pre-screening the sample, I mean estimating propensity scores using DW algorithm (Dehejia & Wahba, 1999, 2002) and restricting estimation to regions of common support. I provide more details on this later.

levels of balance. By different levels of balance, we mean different degrees of the thin or no support problem in different samples that satisfy the DW balancing condition. One may therefore want to further examine the samples in terms of the level of balance they are able to achieve compared to the balance one would expect from a randomized experiment, and rank the control samples according to their ability to replicate the result from a randomized experiment.

3.3 Data description

In this chapter, I explore these arguments by examining the performance of the entropy measure using survey data for which there is an experimental control group and a number of theoretically plausible control groups. The NWS programme data[12] contains treatment and control observations from a RCT, as well as control observations from survey data. In this analysis, the estimate from the RCT (i.e. using the randomized control group) is used as the benchmark. Following Lalonde (1986), plausible control samples from the PSID and CPS are used to calculate treatment effects in a non-experimental setting. The experimental sample includes male respondents in the NSW's ex-addict, ex-offender, and high school dropout target groups. The original Lalonde experimental sample includes 297 treatment and 425 control observations. Dehejia & Wahba (1999, 2002) use a subsample of the Lalonde sample. Their sample has 185 treatment and 260 control observations. This sample is selected to include two years of pre-program earnings which eliminates about 40% of the observations in the original Lalonde sample. It is this ability to control for pre-programme earnings that is the major difference between the two samples. The non-experimental control groups in the CPS and PSID samples contain 15,992 and 2,490 potential control observations respectively. Lalonde (1986) defines plausible control groups that are subsamples of the PSID and CPS data. Table 3.1 shows how these subsamples were selected from the original PSID and CPS datasets. The subsamples can be thought of as crude matched samples of the original control data, or matched on a few variables (gender, age, employment status and income in 1975 and 1976). These represent plausible control groups that a researcher might want to use to estimate the impact of the NSW programme. Even though each of the samples represents a plausible control group, one would expect sample 3 to be a better counterfactual for the NSW

[12] The data is available online from "http://users.nber.org/~rdehejia/data/nswdata2.html".

treatment group than sample 2, and sample 2 in turn to be better than sample 1, because of Ashenfelter's dip (Ashenfelter, 1978).

Table 3.1 Control Samples Description Based on Lalonde (1986)	
PSID-1	All male household heads from data for 1975 through 1978 who are younger than 55 years old and did not classify themselves as retired in 1975
PSID-2	Selects from PSID-1 all men who were not working when surveyed in the Spring of 1976.
PSID-3	Selects from PSID-1 all men who were not working when surveyed in either 1975 or 1976.
CPS-1	All males fulfilling criteria similar to the experimental sample, except those older than 55 years of age
CPS-2	Selects from CPS-1 unemployed males in 1976.
CPS-3	Selects from CPS-1 unemployed males in 1976 whose income in 1975 was below the poverty level.
Experimental	Randomized control sample

This is because control samples 2 and 3 use pre-intevention labour market characteristic to select control units. However, there is no rule about how many years of pre-intervention earnings (or employment history) will be enough to guarantee comparability between the treatment and the control groups. This is where balance comes into play. The literature suggests that based on the way the subsamples are selected, samples that match on pre-programme employment status should be better counterfactuals for the NSW treatment group.[13]

3.4 Method

This study uses the treatment observations from the NWS programme and examines the differences in the treatment effect estimates across control samples. It investigates how these differences relate to balance, as measured by the KS statistic, the standardized difference in means, and the entropy distance metric. Specifically, it focuses on the ability of balance measures to capture variation in the bias and the robustness of treatment effect estimates, across samples and different econometric methods. Results obtained by Lalonde (1986) are not robust across samples because the level of balance achieved in different samples may vary. This is where the metric used in assessing balance can be important. For

[13] For more details on the data, see Lalonde (1986) and Dehejia & Wahba (1999 & 2002).

example, if one control sample has some areas of thin or no support but its mean is similar to the mean of the treatment group, this control sample will be less appropriate than one that exhibits better balance by virtue of having less of a thin or no support problem. Ideally, one would like a situation where balance in the propensity score density after pre-screening the data (i.e. before matching) satisfies

$$f(p(W)|D = 1) = f(p(W)|D = 0) \qquad 3.1$$

Based on the above, the empirical steps taken and discussed in this chapter are as follows:

- Defining the treatment group as the experimental treatment group in the NWS program and 7 plausible control groups (i.e. 3 from each from PSID and CPS data and the experimental control)
- Using the DW algorithm to find the propensity score specifications that balance the covariates for each sample.
- Estimating the KS distance, entropy distance, and standardized difference in means for the propensity score densities of all samples. This is done for both the restricted (pre-screened[14]) and the unrestricted propensity score distributions. Here the propensity score density is used as a proxy for the joint density of covariates, based on Rosenbaum & Rubin (1983).
- Estimating the treatment effect using non-parametric difference in means and other econometric methods.
- Calculating the percentage bias in all samples, compared to the experimental benchmark.
- Calculating the relationship between bias and mismatch measures (entropy, KS, and standardized difference in means). I do this by regressing the bias on the measure, and looking at the strength of the relationship (R^2) and the size of the effect (coefficient).
    - In the case of other econometric methods (i.e. excluding the non-parametric difference in means) this is done by pooling all the bias

[14] By screened sample we mean dropping control whose propensity score lie outside the support. Although one would not normally use such sample for estimation we use it here to illustrate the efficacy or lack thereof balance measures.

estimates (across methods) and their respective balance measures for the regression analysis. This amounts to calculating the average bias for an arbitrarily selected method.

- Calculating the relationship between variability of the estimates (as measured by the standard deviation of the effect estimates across methods) and mismatch measures.

The key question is: Which of the balance measures better predicts bias and variation in effect estimates?

A table showing the propensity score specification for each treatment/control combination is included in Appendix B1 (appendix table A5). All variables that are theoretically relevant are included in the propensity score specification. Following the literature (Dehejia & Wahba, 2002) interaction terms of variables that are not balanced are then included. The DW algorithm is designed to guide the specification of the propensity score equation that balances the sample (i.e. that balances both the covariates and the propensity scores). The algorithm stratifies the sample into blocks based on the propensity scores. It then compares the mean of the propensity scores and the variables included in the propensity score equation within blocks. The appropriate specification is found when the t-test of mean difference across treatment arms are not significantly different within all blocks. Note that the DW algorithm is not designed to equate the propensity score density across treatment arms, as specified by 3.1. Neither is it necessary for it to reject balance based on thin support problems even though it looks at the entire distribution (in blocks). It is merely a way to pre-process the data to deal with obvious problems that may lead to bias. It provides a good way of dealing with violation of common support at the edge of the propensity score density. For example, the Stata implementation of the DW algorithm (Becker & Ichino *et al*, 2002) restricts estimation to common support by dropping control observations that are not within the range of propensity score values in the treatment group. Note that this does not redefine the treatment effect, since treated observations are not dropped. This reduces imbalance caused by some forms of overlap violation. However, there are other problems that this pre-screening does not deal with. Violation of common support can arise when some portions of the support of the treatment observations have no comparable control observations.

Furthermore, in finite samples, there may be areas of thin support i.e. not enough control observations on some portion of the support.

3.5 Results

3.5.1 Results on ranking control samples

Table 3.2 displays the entropy distance (entropy), KS distance (KS), and the standardized difference in means (StD) between the propensity score distributions of the treatment and the control groups. The four panels (A, B, C and D) display the balance measures for restricted (first three rows) and unrestricted (last three rows) samples. By restricted, I mean the sample is restricted to the region of common support based on propensity score values, and the converse is true for the unrestricted sample. The last row in each panel displays the imbalance for the experimental sample. The first thing to note is that the experimental control sample (*Experimental**) has the lowest imbalance for all balance measures. Furthermore, imbalance in the unrestricted samples is always larger than imbalance in the restricted samples.

For example, for the DW PSID sample (Panel A), entropy distance is 0.6058 in the PSID1 restricted sample while it is 0.8109 in the unrestricted sample. The pattern is similar for all other samples. This is expected, since the unrestricted sample contains control observations that don't have any treated observations in their region of support.

In addition, for the distributional measures (KS and entropy), the extent of balance in the distribution of propensity scores varies in a way that agrees with Ashenfelter's dip in all but one sample (the Lalonde PSID unrestricted sample in panel C). In other words, for the entropy and KS measures, control samples that are selected using more information on employment history (recall table 3.1) perform better in terms of balance. For the standardized difference in means, the prediction of Ashenfelter's dip only holds in the DW PSID sample. In Table 3.2, red text indicates the cases where the balance measure does not show a pattern that agrees with what Ashenfelter's dip suggests.

This initial analysis, shows that the standardized difference in means is not the ideal way to rank control distributions in terms of their ability to replicate results from a randomized experiment. This result also does not show any evidence that the distributional measures outperform each other in ranking the control groups. This point is addressed in subsequent sections.

Table 3.2: Entropy distance, KS distance and Standardized difference in means

		DW PSID Sample				DW CPS Sample		
		A				B		
		entropy	**KS**	**StD**		**entropy**	**KS**	**StD**
Restricted	PSID1	0.6058	0.8369	2.0087	CPS1	0.6139	0.7775	**1.3259**
	PSID2	0.2467	0.6181	1.8108	CPS2	0.4902	0.7468	**1.5206**
	PSID3	0.1779	0.5630	1.4210	CPS3	0.2580	0.6137	**1.6497**
Experimental*		0.03	0.20	0.30		0.03	0.20	0.31
Unrestricted	PSID1	0.8109	0.8957	**2.1227**	CPS1	0.8418	0.8724	**1.3956**
	PSID2	0.4263	0.7436	**2.5840**	CPS2	0.6378	0.8229	**1.6298**
	PSID3	0.3758	0.6804	**2.6374**	CPS3	0.3433	0.6885	**1.8880**
Experimental*		0.04	0.21	0.39		0.04	0.21	0.39
		C				D		
		LALONDE PSID Sample				LALONDE CPS Sample		
		entropy	**KS**	**StD**		**entropy**	**KS**	**StD**
Restricted	PSID1	0.4439	0.7599	**1.9735**	CPS1	0.6616	0.8063	**1.4791**
	PSID2	0.3106	0.6584	**2.3109**	CPS2	0.4755	0.7406	**1.7063**
	PSID3	0.2793	0.6116	**2.1200**	CPS3	0.2903	0.6382	**1.8757**
		0.05	0.12	0.21		0.05	0.12	0.21
Unrestricted	PSID1	**0.6425**	**0.8215**	**2.1764**	CPS1	0.7409	0.8369	**1.5015**
	PSID2	**0.3577**	**0.6860**	**2.5551**	CPS2	0.5010	0.7576	**1.7358**
	PSID3	**0.3859**	**0.7037**	**3.0240**	CPS3	0.3088	0.6522	**1.9437**
Experimental*		0.04	0.12	0.20		0.04	0.12	0.20

Red text indicate that the measure did not follow Ashenfelter's dip.

3.5.2 Bias and balance

The previous section shows that distributional measures perform better at ranking controls in the way one would expect, given Ashenfelter's dip[15]. In this section, I examine the relationship between the balance measures and the size of the bias in the treatment effect estimate, compared to the experimental benchmark. Specifically, we examine how correlated and responsive balance measures are to variations in bias. Tables 3.3A and 3.4A show bias as a percentage of the treatment effect estimate from the experimental group in the screened sample. Tables 3.3B and 3.4B contain similar results for the unscreened samples. The treatment effect is estimated with various econometric methods.

[15] What one would expect given Ashenfelter's dip is that balance improves with better selection of control sample.

Columns 1 and 2 (of both tables) display the bias for the unadjusted treatment effect i.e. non-parametric mean difference in outcome and unadjusted difference in difference (Unadj-Diff). Columns 3, 4 and 5 control for covariates using Regression (Regression)[16], difference-in-difference with controls (Diff-in-Diff) and propensity score weighting (pscore weighting)[17]. The specification of the regression model (in column 3, 4 and 5) is based on the specification of the propensity score that achieves balance based on the DW test. Columns 6 to 10 use various matching methods under PSM to estimate the treatment effect i.e. Nearest Neighbour (NN), Radius (Radius), Kernel (Kernel), Stratification (Strat), and Conditional difference-in-difference[18] (CDiD). Columns 11 to 13 contain the entropic distance (S_ρ), the KS distance, and the difference in mean between the propensity score distribution of the treatment and control groups.[19] Column 14 displays the standard deviation of the effect estimates across econometric methods. Standard deviation of effect estimates is used as a measure of variability of the ATT across methods. This is used to assess the robustness of the results across methods.

Under the two hypothesis mentioned in the introduction, the treatment effect from the experimental control group (*treatment effect** row in Tables 3.3A/B and 3.4A/B) is expected to be the least biased and most robust. The lack of bias follows from the strength of randomized experiments, which is confirmed by the balance measures (column 11 to 13). The robustness is shown by the fact that, across econometric methods, the treatment effect using this sample has the lowest variability (column 14). This result also justifies the use of estimates from randomized experiments as the benchmark.

[16] Regression is treated as a cross-sectional estimator. Pre-programme income was therefore not included in its specification, i.e. income in 1974 was not included for the DW sample and income in 1975 was not included for the Lalonde sample.

[17] Following Stuart et al. (2013), treated observations get a weight of 1 while control observations get a weight equal to propensity score over one minus the propensity score. This serves to weight the control group to resemble the treatment group.

[18] Here, the difference-in-difference method is used after nearest neighbour matching.

[19] Note that there is no difference in the rankings of the mean difference in propensity score and the ranking of the standardized difference in means displayed in Table 3.2.

Table 3.3 DW Sample Screen and Unscreened Bias estimates

3.3A: DW Screened and Unscreened samples (% bias)														
	Unadjusted	**Unadj-Diff Regression**		**Diff-in-Diff**	**pscore weighting**	**NN**	**Radius**	**Kernel**	**Strat**	**CDiD**	S_ρ **pscore**	**KS**	**pscore Mean diff**	**Standard deviation of (ATT)**
3.3A: DW Screened														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
PSID1	466	**31**	162	**28**	40	**23**	390	**21**	**7**	**19**	*0.61*	*0.84*	*0.61*	3393
PSID2	155	170	75	69	**4**	**23**	67	43	**27**	**18**	*0.25*	*0.62*	*0.42*	1488
PSID3	49	109	**18**	74	**16**	46	48	57	50	**27**	*0.18*	*0.56*	*0.27*	968
CPS1	226	**10**	40	**2**	**17**	**12**	221	48	**18**	**16**	*0.61*	*0.78*	*0.36*	1675
CPS2	162	**23**	57	46	**19**	**18**	151	**27**	**3**	**13**	*0.49*	*0.75*	*0.47*	1226
CPS3	79	**31**	**35**	**37**	**18**	52	69	**18**	**27**	**27**	*0.26*	*0.61*	*0.42*	646
Treatment effect*	*1854*	*1625*	*1664*	*1664*	*1795*	*2156*	*1905*	*1907*	*1788*	*1786*	*0.03*	*0.20*	*0.03*	*155*
3.3B: DW Unscreened														
PSID1	947	**18**	116	**31**	40	**23**	824	53	**7**	**19**	*0.81*	*0.90*	*0.65*	*6815*
PSID2	303	193	82	52	**7**	**28**	130	44	**27**	**18**	*0.43*	*0.74*	*0.60*	*2344*
PSID3	40	151	85	**19**	**21**	56	55	59	50	**27**	*0.38*	*0.68*	*0.51*	*1258*
CPS1	574	90	68	41	**17**	**13**	543	152	**18**	**16**	*0.84*	*0.87*	*0.38*	*4307*
CPS2	313	56	77	**0.4**	**19**	**18**	279	46	**3**	**13**	*0.64*	*0.82*	*0.50*	*2300*
CPS3	135	60	48	**16**	**18**	52	103	**22**	**27**	**27**	*0.34*	*0.69*	*0.48*	*1041*
Treatment effect*	*1794*	*1806*	*1672*	*1672*	*1799*	*2156*	*1899*	*1897*	*1788*	*1786*	*0.04*	*0.21*	*0.40*	*161*

The table shows percentage bias for ATT using experimental estimate (treatment effect*) as a benchmark for each econometric method.

Percentage bias estimates in bold are for the biases that are within one standard deviation of the experimental estimate (treatment effect*) result.

Columns (1) unadjusted treatment effect estimate; (2) unadjusted difference in difference estimate; (3) regression estimate with controls used for propensity score estimation (with the exception of income in 1974); (4) diff-in-diff with covariate adjustment; (5) propensity score weighting; (6) nearest neighbour matching; (7) radius matching, with radius=0.1; (8) kernel matching (Gaussian kernel was used); (9) stratification matching; (10) conditional difference-in-difference stratification matching technique was used; (11) entropic distance, $\boldsymbol{S_\rho}$**,** between propensity score kernel densities of the treatment and control groups- Gaussian kernel was used for the kernel density estimation; (12) t statistic of test of propensity score means; (13) variance of treatment effect estimates across methods. In each case, the propensity score specification is that which satisfies the mean balancing condition, as in Becker et al. (2002), i.e. mean propensity score and conditioning variables are balanced within each stratum.

Table3.4: Lalonde Sample Screen and Unscreened Bias estimates

3.4A: Lalonde Screened and Unscreened sample (% bias)														
	Unadjusted	**Unadj-Diff**	**Regression**	**Diff-in-Diff**	**pscore weighting**	**NN**	**Radius**	**Kernel**	**Strat**	**CDiD**	S_ρ **pscore**	**KS**	**pscore Mean diff**	**Standard deviation of (ATT)**
3.4A: Lalonde Screened														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
PSID1	920	203	1156	88	262	370	808	299	289	192	*0.44*	*0.76*	*0.54*	*3114*
PSID2	526	134	505	**25**	168	186	247	210	248	116	*0.31*	*0.66*	*0.48*	*1380*
PSID3	132	130	107	118	130	195	168	156	173	211	*0.28*	*0.61*	*0.36*	*302*
CPS1	921	**2**	773	**21**	196	112	880	284	162	144	*0.66*	*0.81*	*0.33*	*3206*
CPS2	518	152	408	78	155	221	467	150	119	91	*0.48*	*0.74*	*0.45*	*1466*
CPS3	220	282	94	206	79	110	143	87	111	206	*0.29*	*0.64*	*0.46*	*613*
Treatment effect*	*902*	*875*	*801*	*845*	*813*	*818*	*921*	*916*	*788*	*903*	0.05	0.12	0.01	*52*
3.4B: Lalonde Unscreened														
PSID1	1858	**50**	1110	154	264	371	1634	390	289	192	*0.64*	*0.82*	*0.59*	*5844*
PSID2	554	**43**	536	**77**	168	174	253	211	248	**116**	*0.36*	*0.69*	*0.53*	*1486*
PSID3	**21**	**71**	164	218	132	195	168	157	173	211	*0.39*	*0.70*	*0.51*	*532*
CPS1	1101	**102**	653	**71**	196	**112**	1050	333	162	144	*0.74*	*0.84*	*0.33*	*3722*
CPS2	573	**84**	393	**127**	155	232	513	157	**119**	**91**	*0.50*	*0.76*	*0.46*	*1612*
CPS3	214	290	**100**	219	**79**	**110**	143	**86**	**111**	206	*0.31*	*0.65*	*0.48*	*610*
Treatment effect*	*886*	*847*	*802*	*802*	*818*	*815*	*912*	*904*	*788*	*903*	*0.04*	*0.12*	*0.01*	*49*

Same as 3.3 above

Recall my hypothesis in the introduction, that, under the assumption that more balance leads to less bias, measures that provide more information about balance should be more correlated with the size of bias than measures that provide less information about balance. I examine this for non-parametric mean difference in outcome, and other econometric methods that control for covariates in the next section.

3.5.3 Balance and bias for non-parametric mean difference

Percentage bias in column 1 of Tables 3.3A & 3.3B and 3.4A and 3.4B show the pattern suggested by the distributional balance measures. Bias in sample 3 is lower than bias in sample 2, which is in turn lower than bias in sample 1. To better summarize this information, I consider how effective each balance measure is in explaining variation in bias across restricted and unrestricted samples. I regress percentage bias estimates on each balance measure. The coefficient, t statistic and the R^2 from the regressions is shown in table 3.5[20].

Table 3.5 Balance and Bias (Non-parametric difference in means)

	Coefficients	t Stat	R^2
DW PSID			
Entropy	856.7819	5.639279	0.864136
KS	502.1405	3.107044	0.65879
StD	152.3375	2.312149	0.516723
DW CPS			
Entropy	505.6841	7.751208	0.923173
KS	346.5839	2.312149	0.782689
StD	149.8906	2.949979	0.6351
Lalonde PSID			
Entropy	1889.343	4.111812	0.771762
KS	1007.805	2.89768	0.62677
StD	258.7576	2.090904	0.466489
Lalonde CPS			
Entropy	1270.155	11.55204	0.963886
KS	834.4531	5.123304	0.839991
StD	323.983	3.21417	0.673861

First, the R^2 is a function of the correlation between bias and balance. Judging by the R^2, table 3.5 shows that the entropy measure is more correlated with the variation in bias than the KS statistic, in all samples. This is because these two measures capture imbalance differently,

[20] For example the first row is the result of running the regression $\%bias = \beta S_\rho + e$

depending on the distributions under consideration. Recall that chapter 2 provides an example where the entropy measure picks up differences in distributions that the KS statistic will ignore. This result therefore suggests that, even though both measures rank control distributions in a similar way, the proposed entropy measure outperforms the KS statistic in terms of quantifying balance. It also support my previous result, that suggests that distributional measures (in general) are better correlated with bias. Second, the coefficients show that the entropy measure is more responsive to changes in bias than the other two measures. This should be expected, given the R^2 values. Therefore, while all balance measures explain significant variation in bias (as shown by the t statistic) their effectiveness in quantifying balance varies, depending on how they asses balance. The important point here is that measures that compare distributions quantify imbalance better than measures that rely on a few moments. Furthermore, while the performance of the measures that compare distributions may be similar (as shown in section 3.5.1 in terms of ranking imbalance), the entropy measure is better at quantifying imbalance. Note that the result presented in this section is based on just 6 observations, it is therefore only indicative of the relationship between balance and bias. The next section investigates a case where there are more observations.

3.5.4 Balance and bias for other econometric methods

In this section, I consider the relationship between balance and bias under other econometric methods. In this analysis, I regress the percentage bias across econometric methods on each balance measure. The results of this analysis therefore show the relationship between balance and bias on average for a randomly selected estimator. Table 3.6 shows the results. Note that the analysis in table 3.6 is based on 54 observations.

The pattern shown in table 3.6 is like the one shown in table 3.5. In other words, for an estimator selected randomly form those considered in this study the proposed entropy measure quantifies balance better than the KS statistic, while the KS statistic quantifies balance better than the standardized difference in means. This is shown by both the R^2 statistics and the coefficients. Again, all balance measures explain statistically significant variation in bias, as shown by the t statistic. Note that the differences between the R^2 values are much lower in table 3.6 than in table 3.5. This is because the results in table 3.6 are based on estimators that implicitly attempt to correct for imbalance in the data (subject to their

assumptions). In other words, bias in the case of non-parametric mean difference reflects the full impact of imbalance in the sample. On the other hand, bias under other econometric methods is mitigated (at least in part) by the implicit correction for imbalance imposed by the method. For example, regression will correct for some imbalance by imputing observations (based on the linearity assumption) in areas where there are support problems. Despite the weaker relationship, the results still suggest that measures of balance differ in the way they capture variations in bias.

Table 3.6 Balance and Bias (other econometric methods)

	Coefficients	t Stat	R^2
DW PSID			
Entropy	153.6606	4.637384	0.288641
KS	102.2266	4.579845	0.283542
StD	33.40873	4.29376	0.258081
DW CPS			
Entropy	101.7508	5.089462	0.328286
KS	73.1092	4.832911	0.305892
StD	33.01117	4.433812	0.270562
Lalonde PSID			
Entropy	698.5967	8.111411	0.553853
KS	396.92	7.431922	0.510318
StD	108.3511	6.31466	0.42934
Lalonde CPS			
Entropy	433.0008	8.361686	0.568818
KS	299.8788	8.06255	0.550865
StD	122.7897	7.157537	0.491511

The relationship between balance and bias also has implications for variability in treatment effect estimates across various econometric methods. When a sample satisfies equation 3.1 or is close enough to satisfying equation 3.1 there will be little need for any form of adjustment imposed by econometric methods. Under this condition of balance, weights imposed by matching (or any other method) will have little influence on the inference. However, this is often not the case in observational studies. For example, in the case of matching estimators, matching methods are used to correct for imbalance in the pre-screened sample so that the (weighted) matched sample satisfies the equality of the propensity score density described by 3.1. The problem is that matching methods differ in the weight they attach to each observation in an attempt to satisfy 3.1. This is because various

matching schemes used for PSM correspond to different weighting functions. In large samples, all matching estimators should yield the same result (Smith, 2000). However, different weighting functions often give different results in small samples. This difference in weighting can lead to inferences that are not robust, especially when the level of imbalance in the pre-screened sample is high. Results in this case will be influenced by weights imposed by the matching method, and consequently be model-dependent.

A similar argument can be made when estimation is by regression method. Angrist and Pischke (2008) compare PSM and regression methods, and show that regression is equivalent to using a different weighting scheme than the one used under PSM. PSM weights the covariate-specific estimate into an estimate of the ATT, using the distribution among the treated units. Regression, however, produces a variance-weighted average of these effects (Angrist and Pischke, 2008: 54). In the next section, I explore the relationship between imbalance and variability of effect estimates across methods as implied in our second hypothesis.

3.5.4 Robustness of treatment effect across econometric methods

The standard deviation of effect estimates across methods is shown in column 14 of tables 3.3A, 3.3B, 3.4A and 3.4B, and is used as a measure of the variability of effect estimates. Under our second hypothesis, we expect good balance measures to predict variability of effect estimates across methods better than other balance measures. This is based on the idea that imbalance introduces lack of robustness of treatment effect, as explained in the last section. To make the explanation clearer consider a situation where there is support problem on some portion of the support and a regression method is used to estimate the ATT. Since regression assumes that the relationship between the outcome and covariates is linear, this method will impute values for the missing observations based on the linearity assumption, and this will be reflected in the result. If nearest neighbour matching is used instead, the missing observations will be replaced by the observation nearest to them. Consequently, estimates from nearest neighbour matching and regression may yield very different results. In a situation where balance in the pre-screened sample satisfies equation 3.1, this difference in econometric methods will not have a huge effect on the results across methods. This is because there is no imbalance to be corrected, so the method used to calculate treatment effect will not matter. The point is that if the balance measure captures balance well it should also predict

the influence of the weights imposed by various econometric methods which will be captured by the variability of effect estimates.

Column 14 shows that sample 3 yields treatment effect estimates with lower variability across methods than sample 2, and sample 2, in turn yields results with lower variability than sample 1. I summarize the relationship between balance and variability by considering the correlation between balance measures and our measure of variability of effect estimates. Table 3.7 shows the results. Note that the results shown in this table are also based on 6 observations.

Table 3.7: Bias and Variability of effect estimates across methods

	Entropy	**KS**	**StD**
DW PSID	0.9399**	0.8928**	0.1197
DW CPS	0.9261**	0.8729*	-0.5101
Lalonde PSID	0.9383**	0.9154**	-0.4487
Lalonde CPS	0.9850**	0.9559**	-0.9672**

Distributional measures (entropy and KS) again outperform the measure based on comparing a few moments (StD). Furthermore, the proposed entropy measure is more correlated with variability of effect estimates across methods than the KS statistic. The KS distance also performs well, but its correlation is weaker than the one observed for the entropy measure. Like the result on predicting bias, the standardized difference in mean is shows the worst performance.

This result suggests that, while econometric techniques can be used to estimate an unbiased treatment effect (giving the assumption(s) of the method), the entropic metric can help identify samples that yield more robust results across econometric techniques. Such samples should be preferred because their results rely less heavily on the model's assumption(s).

3.5.5 General discussion of the results

Another way to explain the results discussed in this chapter is to think in terms of outliers. Note that all samples contain a set of observations that are common to all samples. For example, CPS 1 and 2 both contain observations contained in CPS 3. This is because CPS 2 is a subsample of CPS 1 and CPS 3 is a subsample of CPS 2 (see table 3.1). Therefore CPS 1 and 2 can be thought of as containing more "outliers", i.e. units that are more dissimilar to units in the treatment group than to CPS 3. This may affect their ability to balance the covariate

distribution of observations in the treatment group, even though all samples satisfy the DW algorithm. A plausible explanation for this is that the propensity score is not a permanent tag for each observation (Lee, 2013). Therefore, the presence of these "outliers" may change the propensity score value attached to observations in the (core) CPS 3 sample across the three control groups. This, in turn, may affect the matched sample, since this depends on the propensity score value.

Evidently, sample 3 should be preferred, because it shows the least bias under the adjusted mean difference approach, and is most robust across methods. However, data that show a level of balance like data from a randomized experiment is often not available. In such situations (or even where data exist that show a reasonable level of balance), the effect of the suitable econometric approach cannot be dismissed. This is because, when the assumptions that validate these methods are right for the data, these methods can produce very good results. Heckman, Ichimura, Smith and Todd (1998) note that both data and method matter in estimating unbiased treatment effects. I am therefore not advocating against the use of non-experimental methods. On the contrary, my main point is that results from non-experimental studies can be improved if a control sample with covariate distributions that are closer (in entropic sense) to those of the treatment group is used (in a case where there are alternatives). It is obvious that samples that exhibit better balance should be preferred. The importance of this study is in identifying a way to get at the sample that achieves the best balance possible.

The proposed measures for assessing balance in different control samples should be used with caution. The entropic measure can only measure observable differences in the distribution of covariates. Thus comparison across data sources or samples with different sets of covariates may be problematic. For example, the DW sample controls for 1974 income, while the Lalonde sample does not. It is not clear how this variable will affect the entropic distance of the propensity score density in the two samples. One might expect that, since the Lalonde sample does not have to control for imbalance in this variable, it should have the lower entropic distance than the corresponding DW sample. However, the results do not always portray this.

In general, using the entropic distance metric to rank control distributions that are not identical in terms of conditioning variables or survey instruments might yield dubious results.

When comparing control samples across datasets, it is not guaranteed that the set of conditioning variables will be identical. Even when they are, differences in survey instruments may mean that samples are different in terms of unobservables.

Another relevant issue is how to interpret the differences between the entropic distances associated with different samples. This analysis suggests that a control group whose propensity score distribution has smaller entropic distance from the treatment group performs better than a control sample with larger entropic distance from the same treatment group. Such control samples are more likely to recover an unbiased treatment effect or a treatment effect estimate that is comparable to one obtainable under a RCT. However, this analysis cannot reliably pin down how different entropic distances must be for there to be a significant difference between the level of bias in different samples .

3.6 Conclusion

This chapter examines the plausibility of using the entropic distance, the KS statistic, and the standardized difference in mean to rank the ability of non-experimental control distributions to replicate experimental results. Relying on the result of Rosenbaum & Rubin (2009), the propensity score density is used to assess balance in the multivariate distribution of covariates. The results show that, in non-experimental situations where one can define plausible control groups with an identical set of covariates, the entropic distance measure performs better at identifying control groups that provide estimates comparable to those from a RCT.

The result also suggests that relying on the DW algorithm alone might result in a situation where information that can affect inference is ignored, especially when there are competing control samples. It is shown that balance (as measured by the proposed entropic measure) in the pre-screened sample can predict how successful estimation will be in mitigating bias and variability across different methods. More generally, there can be cases where imbalance will not be reflected in measures of balance that compare mean and variance. This imbalance does matter, and will show up better in distributional measures. Of the distributional measures considered in this study, the entropy measure performed better than the KS distance measure in capturing balance.

4 *Estimating the Impact of the South African Child Support Grant using a Genetic Algorithm and entropy measure*

4.1 Introduction

The previous chapter considers a case where we have experimental data. However, researchers often do not have access to an experimental control group. Non-experimental data is therefore used to recover an estimate of the ATT. As noted above, when a control sample that has large entropic distance from the treatment sample is used, ATT is more likely to be biased. One way to mitigate bias is to use methods that optimize balance. This can be thought of as matching (or weighting) until balance cannot be improved in all covariates. The analysis discussed in this chapter makes use of a method that optimizes balance. The method introduced by Sekhon & Diamond (2005) is used to estimate the impact of the Child Support Grant (CSG) in South Africa. This algorithm is called Genetic Matching (GenMatch), it performs multivariate matching by using an evolutionary search algorithm to determine the weight each covariate is given, to optimize balance in the matched data. The algorithm allows the researcher to select the preferred measure of balance, and a fitness function to be optimized. The default balance measures used by the algorithm are the standardized difference in means and the KS statistic. This chapter illustrates how the entropic distance measure can be used with this matching algorithm, and compares its performance with that of other balance measures.

The research in this chapter investigates the CSG programme, which is one of the social assistance programmes of the South African Government, targeted at children. This program is maintained at significant cost to South African taxpayers, so there have been several attempts to estimate its impact. The analysis in this chapter follows the work of Coetzee (2011 & 2013 &)[21], which makes use of data from wave 1 of the National Income Dynamics Study (NIDS) for 2008 to estimate the impact of CSG on six outcome variables that capture child well-being. However, the application uses only one outcome variable (which is the height-for-age z score). Under the assumption that the treatment variable is binary (i.e. a child is either receiving a CSG or not), Coetzee (2011) finds no convincing evidence that the CSG results in

[21] Coetzee (2011) is the working paper version of Coetzee (2013).

improvement in well-being in terms of health, education, and household expenditure (using PSM methods).

In this chapter, the treatment effect of CSG on height-for-age z score is re-estimated in the binary case. The PSM method used by Coetzee (2011) is replaced with an approach that is expected to improve balance in the matched sample (i.e. GenMatch). This analysis illustrates one way in which the entropy measure can be used as an alternative measure of balance. It also shows that the way balance is measured matters for analysis outcomes. Other things being equal, the results under the different measures of balance (i.e. entropy, KS, and Standardized difference in means) differ. The treatment effects are 9%, 15% and 16% of standard deviation, respectively, when the standardized difference in means, KS statistic, and the entropy measures are used as balance measures[22]. The difference in effect estimates can only be attributed to difference in the way balance measures capture imbalance. We note that while the standardized difference in means will choose optimal weights that attempt to balance only the mean and the variance of the covariates, the entropy measure and the KS statistic will choose optimal weights that attempt to balance the distribution of covariates. Furthermore, since the entropy measure captures imbalance in distribution in a way that is different from the way that the KS statistic captures imbalance, their outcomes too will be different. As a consequence, the distribution of weights allocated to variables under the different balance measures are different. This in turn affects the inferences from the analyses.

4.2 Literature Review

4.2.1 Review on genetic matching

GenMatch seeks to maximize covariate balance by finding optimal covariate weights. This is achieved by optimizing a user-specified fitness function, which is in turn a function of some balance measure. For example, one can choose to maximize the mean of the p-values of t-tests for all covariates. In this example, the mean is the balance measure, while the fitness function is some function of "mean of p-values". Alternatively, one can choose to maximize the minimum p-value of t-tests for all covariates. In general, the aim is to optimize balance as much as possible rather than using a stopping rule (i.e. critical value in a statistical test). Diamond & Sekhon (2013) argue that this method will help address some limitations of

[22] Note that, given our results discussed in the previous chapter, it is not surprising that the performances of the KS statistic and the entropy measure are similar.

popular matching procedures, such as the Mahalanobis distance and propensity score matching. According to the authors, the problem with these methods is that they may make balance worse in some covariates in finite samples. This is because these methods are not equal percentage bias reducing (EPBR) when covariates (W) have distributions that are not ellipsoidal, such as normal or t-distributions (Diamond and Sekhon, 2013). A matching method is EPBR for covariates when the percentage reduction in the biases of each covariate is the same (Diamond & Sekhon, 2013). When matching is not EPBR, bias for some linear combination of elements of W (the covariates) is increased, even when covariate means are closer in the matched data than in the unmatched data (Rubin, 1976a). In this regard, Genmatch can be thought of as a generalization of the Mahalanobis metric to include an additional weight matrix:

$$d(w_i, w_j) = \left\{ (w_i - w_j)' (S^{-1/2})' \, M \, S^{-\frac{1}{2}} (w_i - w_j) \right\}^{1/2} \ldots\ldots\ldots 4.1$$

where M is a $t \; x \; t$ positive definite weight matrix, S is the variance co-variance matrix of W and $S^{\frac{1}{2}}$ is the Cholesky decomposition of S. The main goal is to find the weight matrix M that achieves the best balance when the distance produced by $d(w_i, w_j)$ is used to match observations in the sample. Propensity scores can be included as one of the covariates. In this case, both propensity score and Mahalanobis matching can be thought of as limiting cases of GenMatch. If propensity scores contain all relevant information in the covariates, then all other variables will receive a zero weight. In this case, GenMatch is equivalent to PSM. On the other hand, GenMatch will converge to Mahalanobis distance (even when propensity scores are included) if it is the more appropriate distance measure for the sample (i.e. when the propensity scores fail to achieve the best level of balance in the covariates). In less extreme cases, GenMatch allocates weight to propensity scores and all covariates. The implication is that it does not only balance the propensity scores, but also accounts for imbalances in individual covariates, where the imbalance is not accounted for by the propensity scores.

By default, GenMatch optimizes (maximizes) the p-values of t-tests and KS tests. In my application I use the entropy measure as the balance metric to be optimized. Therefore, instead of maximizing the minimum p-values of the t-test and KS test, I minimize the minimum entropic distance between the distributions of covariates. One could optimize the p-value of

the entropic distance too, but I do not explore that in this study, in order to be consistent with how the entropic distance has been used in my analysis covered in previous chapters.

GenMatch searches for the best balance possible by generating random solutions, i.e. it generates a number of random weight matrices M[23] . These solutions are then used to estimate equation 4.1, and for each solution, balance is checked in the matched sample produced by using the distance defined in 4.1. Note that in this analysis 1:1 matching with replacement is used. A solution that arises from weight matrix M_i is preferred to another solution M_j, if M_i produces more balance in the matched sample according to the fitness function supplied by the user[24]. The default function in GenMatch (which is used in this study) sorts all balance statistics from the most discrepant to the least. The random solutions are assessed by their ability to minimize the maximum discrepancy. If multiple sets of weights (M) result in the same maximum discrepancy, the second largest discrepancy is examined to choose the best weight. This process continues iteratively until all ties are broken (Sekhon, 2011). After assessing balance and ranking the solutions in the first population according to their fitness values, a new population of solutions is formed. This is done using genetic operations: Mutation, crossover, and selection. These operators work on one or more current trial solutions from the current population to produce one or more trial solutions in the new population[25]. The new population is then assessed and ranked, using its fitness values. This process continues until the balance statistics can no longer be improved.

4.2.2 Brief review of the South African literature on the Child Support Grant (CSG)

The CSG is one of the social assistance programmes of the South African Government which is targeted at children. Introduced in 1998, the CSG is an unconditional grant intended to assist poor households in improving the welfare of children in such households. While there are soft conditions related to school attendance attached to the grant, failure to comply is not exclusionary. The unconditionality of the grant therefore gives full financial autonomy to

[23] This number is called population size in Genetic Algorithms. For our analysis, the population size is 2000.
[24] As noted earlier, this fitness function can be to minimize the mean of the balance statistics across all covariates.
[25] Selection gives preference to better the solution to make it into the next generation of solutions (or the offspring population). Crossover combines two or more current solutions to form a new solution (offspring in the new population). Mutation is used to encourage diversity amongst solutions. This is achieved by changing parts of a candidate solution in the current population randomly to produce new solutions. See Mabane and Sekhon (2011) for more details.

caregivers to spend the grant. Consequently, measuring the impact of the CSG is more complex than measuring the impact of conditional cash transfer programmes such as *Oportunidades* in Mexico. The Mexican grant is conditional on school attendance, health (in the form of clinic visits) and nutrition assessments. For this programme, the direct response variables would be school progress, health status and nutrition status. By contrast, the CSG has a multitude of potential response variables. This makes estimating the impact of CSG a difficult prospect. This programme is costly to South African taxpayers (approximately 35.5 billion in 2011, according to South African National Treasury. Consequently, several studies have investigated the impact of the programme on child welfare.

Outcome variables that have been explored in these studies include variables that capture health, nutrition, and education. Case, *et al.* (2005), using the KwaZulu Natal Income Dynamics Study 1993-2004, find that children who benefit from the CSG are more likely to be enrolled in school, than older siblings who did not benefit from the CSG. Aguero, *et al.* (2006), using the same dataset, and under the assumption that treatment effect is continuous, find that a high dosage of CSG early in life has a positive impact on a child's nutritional status. Coetzee (2013) investigates the impact of CSG on a few welfare variables (health, education, and nutrition) using data from wave 1 (2008) of the National Income Dynamics Study. Like Aguero, *et al.* (2006), Coetzee (2013) finds an effect under the assumption that treatment effect is continuous. However, the results presented in Coetzee (2011) show that, in the binary case, the CSG had no significant effect on any of the outcome variables concerned with child welfare. A significant negative effect was found only for the adult expenditure variable in her study. In the continuous treatment case, treatment effect depends on the length of time the care-giver has received the CSG for the child. The binary case assumes an equal effect on children who have benefited for longer periods of their lives and those who have benefited for a shorter period. Duration of receipt is expected to have a differential impact on the effect of the CSG. Treating these two groups the same is therefore likely to result in underestimation of the effects of the programme (Aguero, *et al.,* 2006). As a result, the binary case serves to estimate the lower bound of the treatment effect.

4.3 Methods

The entropic distance is used as a balance measure to be optimized in estimating the effect of the CSG. Its performance is then compared with the performance of the standardized

difference in means and the KS statistic. I use the DW algorithm to estimate the propensity scores, and include it with the covariates GenMatch seeks to balance. Aside from using a different balance measure (entropy metric) and a different data processing method (GenMatch instead of PSM), a number of changes are made in this analysis that differentiate it from the analysis of Coetzee (2011). First, the treatment group is redefined to include children whose care-givers have received the CSG for at least 34% (a third) of their lives. This choice was made to mitigate the effect of treatment group members that dilute the effect of the CSG. This is because the treatment effect for treatment group members receiving the CSG for less than 34% of their lives is likely to pull the average effect down in the binary treatment case[26]. This restriction also helps in finding a propensity score specification that balances the covariates (according to the DW algorithm). Even though it is not compulsory to include propensity scores in the set of covariates given to GenMatch, including propensity scores that balance the covariates reduces computation time for GenMatch (in a case where the propensity scores are informative about balance).

Second, Aguero, *et al.* (2006) notes that (unobserved) caregiver motivation can bias the treatment effect estimate. A different approach to the one used by Coetzee (2011) is used to recover the unobserved motivation. Caregiver motivation is directly related to the length of time that a caregiver takes to apply for the CSG, with highly motivated caregivers applying earlier than less motivated caregivers. A delay in applying for the programme affects the impact via the length of time the child benefits from CSG. Conversely, early application increases the duration of grant receipt, or treatment dosage. For example, receipt of the CSG can improve nutrition through increased spending on food. However, nutritional deficiencies in early stages of life can cause stunting (measured by height-for-age) which may not be reversible (Duflo, 2003).

Aguero, *et al.* (2006) therefore constructs a variable that captures variation in the motivation of care-givers as a function of the amount of time that passes before each caregiver applies for the grant, and whether they reside in a rural or urban area. It was argued that motivation is a function of effectiveness of CSG rollout in the area where the caregiver lives. Coetzee

[26] This is influenced in part by the result of Coetzee (2013).

(2013) notes that even though the programme was rolled out simultaneously in all areas, delay in uptake is much shorter for rural areas than for urban areas.

4.3.1 Caregiver motivation – Coetzee's (2011) approach

Coetzee (2011) calculated a variable that captures caregiver motivation as the difference between actual and expected delay, given the age of the child, and location (rural or urban). Expected delay was estimated using OLS for children born two years or more before the NIDS 2008 survey. This is done because average delay for children under two years will be underestimated. Many eligible children in this age cohort are not yet benefiting from the CGS (Aguero, *et al.*, 2006 & Case, *et al.,* 2005). First, actual delay is calculated for each child who is eligible for the grant. For those receiving the grant, delay is calculated as the number of days between the birth date of the child and the date the CSG was first received for the child. For non-recipients, the delay is calculated as the number of days between the child's birth date and the date of the interview. The expected delay is then calculated as the OLS prediction of the delay as a function of the child's age and location. The difference between the actual and expected delay is then standardized[27] to arrive at a variable that represents the unobserved variation in caregiver's motivation to apply for CSG. The resulting variable is thus, by construction, a strong predictor of treatment.

4.3.2 Caregiver motivation censored regression approach

The approach adopted in this research to calculate caregiver motivation is different from the one used in Coetzee (2011). By construction, the approach discussed in the previous section guarantees that there will be imbalance in the data, since the motivation variable must be different across the treatment and control groups. This does not necessarily imply that respondents in the two groups are different. It is more likely the consequence of not accounting for the fact that delay in the control group may not be equal to the child's age. Some caregivers in the control sample may have applied for the CSG, and not received it.

The point of departure is therefore in the definition of observed delay for the eligible non-beneficiaries. In the approach just described, observed delay for control observations is the difference between the birth date of the child and the date of the interview. The data, however, contains information detailing whether or not the caregiver ever applied for CSG on

[27] Calculated as delay minus mean of delay over standard deviation of delay.

behalf of the child, and a subsequent question asks about the date of application. To account for this I construct a control sample that uses the date of application (instead of interview date) for eligible non-beneficiaries, to calculate the delay for the respondents that supply this information (I call this control sample 1). Although this reduces the sample size in the control group considerably, compared to the approach adopted by Coetzee (2011), this control sample is more appropriate. Furthermore, the reasons why these caregivers were unsuccessful in their application may not be related to the treatment. Table 4.1 shows the responses from caregivers when they were asked why they had not applied for the CSG for the child in their care.

Table 4.1 Reason why CSG has not been applied for

1. Caregiver has not heard of CSG
2. Caregiver does not know how to apply for CSG
3. CSG applied for by someone in another household
4. Ineligible because child is too old
5. Caregiver cannot apply as not child's mother
6. Child is not eligible as receives a different grant (foster care/care dependency)
7. Child is not eligible as caregiver income too high
8. Caregiver does not have the right documentation (e.g. Birth certificate, ID
9. Cost of application is too high
10. Application process is too complicated or too time-consuming
11. In process of applying or getting relevant documentation
12. Haven't got round to it yet
13. Cannot be bothered
14. Other
15. No need
16. Parent/s work for government
17. Parent/s working

To improve on the sample size, a second control sample is constructed. This control sample is made up of control sample 1 and control units that are eligible non-beneficiaries but have never applied for the CSG (see table 4.1). The reasons for not applying can be broadly divided into three categories: The first category includes observations with responses 1, 2, 4, 5, 7, 9, 16 and 17, shown in table 4.1. These are children whose caregivers appear not to have applied because of lack of awareness of how the CSG works. The second category are observations with responses 10, 11, 12 and 13, shown in table 4.1. These are caregivers who appear not to be interested in applying for CSG. The third category are observations with responses 3, 8, 14 and 15, or where no reason was given for not applying. To accommodate these observations,

censored regression is employed to estimate the expected delay equation. These observation are regarded as being right censored, with a variable censoring point that is equal to the age of the child (since application has not been done for these children, but may be done sometime in the future).

4.4 Data and summary statistics

The first wave of the National Income Dynamics Study (NIDS) is used for the analysis. The NIDS dataset is a nationally representative panel dataset, begun in 2008. Table 4.2 presents the summary statistics of the variables used in this analysis. The treatment group consists of children under the age of 14 (which was the age limit in 2008). Eligibility for CSG is determined by age and a means test[28] as it was applied in 2008. The age and means test condition is used to identify children whose care-givers should be able to receive the CSG. A child is assigned to the treatment group if it is indicated that CSG is currently being received for the child[29]. The control group consist of children who were eligible but have never benefited from the CSG. This classification and subsequent cleaning of the data were undertaken to follow Coetzee (2011) as closely as possible.

Furthermore, in a similar manner to Coetzee (2011), the treatment group is separated into three categories. These categories reflect the varying length of time for which children in the sample have received the CSG. This is important because of the assumption that the grant receipt period should be correlated with the effect of the grant (Aguero, *et al.*, 2006). Columns 1, 2 and 3 of table 4.2 show the different categories. Low (dosage) refers to children who participated in the program for 0–34% of their life. Medium (dosage) refers to children who participated for 34–67% of their life. Finally, high (dosage) refers to those children whose caregivers received the grant for 67–100% of the child's life. As noted earlier, the treatment sample is defined as children who receive medium or high dosages. The 4th column shows a summary of all eligible treated children. Columns 5 and 6 show the summary statistics of eligible and non-eligible control group members. Column 7 shows the difference in means of

[28] At the time of the NIDS 2008 survey, a caregiver (who does not have to be a family member of the child) must have a monthly income below R800 in urban areas or R1,100 in rural areas.

[29] It should be noted that there are children whose care-givers received the CSG for them in the past , but who are not presently receiving this grant. We exclude these respondents because of the possibility that they may dilute the effect of the CGS since we cannot be certain of how long they received the CSG.

characteristics across treatment status for the treatment and eligible control groups, while column 8 shows the same statistics for the treatment and ineligible control groups.

While there are significant differences between the treatment units and the eligible and ineligible control units, the mean differences between the treatment and the ineligible control groups are larger than corresponding values that compare the treatment and eligible control groups.[30] The different treatment categories (low, medium and high) are more similar to one another than to the control groups. The only exception is the caregiver motivation variable.[31] As expected, the summary statistics show that the length of time the child has benefited from CSG increases with increased caregiver motivation. On average caregivers of children who enjoy high dosage of CSG delay for a year while those that are in the low dosage category delay for about eight years. Similar patterns are observed by Coetzee (2011). In terms of univariate comparison, since most of the variables are dummy variables, comparing the proportion will suffice rather than using the entropic measure, since both measures should be equivalent.

Table 4.3 shows the descriptive statistic for the redefined treatment group (i.e. treatment dosage of at least 34%) and the control samples 1 and 2 (last two columns of table 4.3). Table 4.4 shows the descriptive statistics for the motivation variable in each combined sample. As mentioned earlier, the main problem with the approach described in section 4.4.2 is that it reduces the sample size considerably. Table 4.4 shows that there are 107 observations in control sample 1 while control sample 2 has 366 observations. There are 2,492 observations in the treatment group. This suggests that the standard error of the estimate may be high, because matching will require using each control observation multiple times.

[30] Note that the race group white is excluded. This is because the number of observations in that category is negligible compared to other groups.

[31] The motivation variable shown in table 4.2 is calculated using the approach in Coetzee (2011), i.e. the interview date instead of the application date was used to calculate delay.

Table 4.2 Summary Statistics by treatment /Control Categories														
	Low		Medium		High		All Treated units		Eligible Control		Ineligible Control		Treatment vs Eligible Control	Treatment vs Ineligible Control
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	Mean diff	Mean diff
Caregiver Characteristics														
Motivation	-1.10	0.44	0.06	0.47	0.84	0.72	0.05	0.98	-0.80	0.93	.	.	-0.85***	
delay_yr	7.60	3.38	4.11	2.05	1.06	0.82	3.87	3.48	6.52	4.08	.	.	2.65***	
Employed	0.26	0.44	0.29	0.45	0.25	0.43	0.27	0.44	0.27	0.44	0.93	0.26	0.003	0.66***
married	0.56	0.50	0.56	0.50	0.51	0.50	0.54	0.50	0.72	0.45	0.72	0.45	0.18***	0.18***
Education	11.26	8.17	10.75	7.97	11.07	6.31	11.03	7.40	11.25	7.00	12.82	6.22	0.22	1.79***
Age	40.69	13.12	38.31	12.06	36.47	11.87	38.24	12.42	41.58	13.83	39.72	9.13	3.34***	1.48**
Child Characteristics														
Age	8.66	3.92	7.78	3.66	6.08	2.95	7.34	3.64	8.50	4.08	8.71	3.85	1.16***	1.37***
Gender	0.49	0.50	0.49	0.50	0.50	0.50	0.49	0.50	0.51	0.50	0.54	0.50	0.021	0.05*
Black	0.88	0.32	0.89	0.31	0.89	0.31	0.89	0.31	0.73	0.44	0.60	0.49	-0.16***	-0.29***
Coloured	0.11	0.31	0.10	0.30	0.10	0.31	0.11	0.31	0.25	0.43	0.35	0.48	0.14***	0.24***
Asian	0.01	0.08	0.00	0.05	0.00	0.07	0.00	0.07	0.02	0.15	0.05	0.22	0.02***	0.05***
HH Characteristics														
Electricity	0.67	0.47	0.66	0.47	0.69	0.46	0.68	0.47	0.75	0.43	0.87	0.34	0.08***	0.19***
Water	0.22	0.55	0.19	0.48	0.21	0.47	0.21	0.50	0.40	0.52	0.69	0.50	0.19***	0.49***
Telephone	0.05	0.22	0.03	0.17	0.03	0.18	0.04	0.19	0.11	0.32	0.25	0.43	0.08***	0.21***
Toilet	0.29	0.45	0.28	0.45	0.32	0.47	0.30	0.46	0.47	0.50	0.83	0.38	0.17***	0.53***
HH head Gender	0.54	0.50	0.57	0.50	0.55	0.50	0.55	0.50	0.67	0.47	0.65	0.48	0.12***	0.01***
HH log pa capita Expenditure	5.52	0.70	5.55	0.70	5.58	0.70	5.55	0.70	5.92	0.94	6.97	1.03	0.37***	1.42***

Table 4.3 Summary Statistics treatment group / Control Samples								
	Treatment (medium/High dosage)		CONTROLS				t test means Treatment vs controls	
			Sample 1		Sample 2			
	Mean	SD	Mean	SD	Mean	SD	1	2
Caregiver Characteristics								
delay_yr	2.36	2.11	6.52	4.08				
Employed	0.27	0.44	0.33	0.47	0.28	0.45	0.07	-0.00
married	0.53	0.50	0.75	0.44	0.72	0.45	0.22***	0.20***
Education	10.93	7.06	9.53	6.29	11.77	7.65	-1.40*	0.51
Age	37.25	11.98	40.17	13.30	41.87	14.68	2.91*	3.95***
Child Characteristics								
Age	6.81	3.38	8.13	3.99	7.96	4.26	1.32***	1.63***
Gender	0.50	0.50	0.52	0.50	0.54	0.50	0.02	0.02
Black	0.89	0.31	0.77	0.43	0.77	0.42	-0.12***	-0.18***
Coloured	0.10	0.30	0.23	0.43	0.20	0.40	0.13***	0.15***
Asian	0.00	0.06	0.00	0.00	0.01	0.07	-0.003	0.03***
HH Characteristics								
Electricity	0.68	0.46	0.69	0.46	0.66	0.48	0.13**	0.01
Water	0.20	0.47	0.34	0.58	0.33	0.50	0.006	0.21***
Telephone	0.03	0.18	0.04	0.19	0.07	0.26	0.08	0.08***
Toilet	0.30	0.17	0.39	0.49	0.39	0.49	0.09	0.18***
HH head Gender	0.55	0.50	0.65	0.48	0.69	0.46	0.13*	0.13***
HH log pa capita Expenditure	5.56	0.70	5.70	0.78	5.69	1.01	0.06	0.39***

***Number of observations in each category is shown in appendix C1.**

Table 4.4 Summary of caregiver motivation for treatment and various control samples				
			Controls	
		Treatment	1	2
Sample 1	Mean	0.035	-0.819	
	SD	0.988	0.936	
Sample 2	Mean	0.404	-0.251	-0.992
	SD	0.912	0.775	0.289
N		2492	107	366

4.5 Results

The analysis is performed for control samples 1 and 2. Note that the difference between the two control samples is that control sample 1 contains control units with the date of application while control sample 2 is made up of units in control sample 1 and other units whose motivation variable was recovered by censored regression because there is no information on the application date. There are no major differences in the findings from the two samples. Four different analyses are undertaken to tease out the effect of the changes I made to the treatment/control sample, and the effect of the balance measure used with GenMatch to calculate the treatment effect. The logistic probability model is used in the propensity score equation. For control sample 1 the specification that satisfy the DW balancing condition is given by

$$\begin{aligned} Pr\,(CSG = 1) = {} & \beta_1 + \beta_2 employed + \beta_3 married + \beta_4 electricity + \beta_5 water + \beta_6 telephone \\ & + \beta_7 hh\ head\ gender + \beta_8 caregiver\ years\ of\ education + \beta_9 caregiver\ age \\ & + \beta_{10} colured + \beta_{11} asian + \beta_{12} black + \beta_{13} motivation + \beta_{14}\ motivation \\ & * employed + \beta_{15} employed * hh\ head\ gen + \beta_{16}\ caregiver\ age \\ & * caregiver\ years\ of\ education \end{aligned}$$

For control sample 2 (with the same treatment sample) the specification is given by

$$\begin{aligned} Pr\,(CSG = 1) = {} & \beta_1 + \beta_2 employed + \beta_3 married + \beta_4 electricity + \beta_5 water + \beta_6 telephone \\ & + \beta_7 hh\ head\ gender + \beta_8 caregiver\ years\ of\ education + \beta_9 caregiver\ age \\ & + \beta_{10} colured + \beta_{11} asian + \beta_{12} black + \beta_{13} motivation + \beta_{14}\ motivation * \\ & * motivation + \beta_{15} motivation * employed \\ & + \beta_{16}\ caregiver\ years\ of\ education * married + \beta_{17}\ hh\ head\ gender \\ & * married \end{aligned}$$

The samples are then restricted to the region of common support. As in the the analysis discussed in the previous chapter, this was done using the common support option of the "pscore" command in Stata (see Becker & Ichino *et al*, (2002)).

Note that finding a specification that balances the covariates according to the DW algorithm is not necessary for Genmatch. In other words, one could use any propensity score specification irrespective of whether it achieves balance according to DW conditions (by this, I mean one could, for example, use a specification that excludes all higher order terms). The important thing for GenMatch is for the propensity score to contain some information about balance. However, I use the one that satisfies the DW balancing condition so that I can comment about the multivariate balance in the sample I use in a way that is consistent with the previous chapter. The entropic distance between the propensity score distribution of the treatment and control groups is 0.40 and 0.65 for control samples 1 and 2 respectively. As mentioned in the previous chapter, both samples need further imbalance correction so that one should expect the treatment effect in these samples to vary across econometric methods. I also note that sample 2 has higher imbalance than sample 1.

4.5.1 Analysis 1: Propensity Score Matching

This section presents the result of using PSM on our redefined sample. The treatment effect, as shown in table 4.5, is 9% and 7% of standard deviation respectively, when control samples 1 and 2 are used with the treatment group (Note that the estimate in Coetzee (2011) is 7% of standard deviation). The conclusion from this analysis is not very different from the one in Coetzee (2011), in that the treatment effect is not statistically significant. The size of the effect is slightly larger when control sample 1 is used; this may be attributed to the use of application date data rather than the birth date of the child to calculate motivation.[32]

Table 4.5 Treatment effect estimate and Standard errors for PSM

	Sample 1	Sample 2
Estimate	0.0945	0.07389
SE	0.3159	0.33067
T-stat	0.3000	0.2200

[32] The treatment effect is calculated with the "psmatch2" command in Stata (see Leuven and Sianesi (2003)).

4.5.2 Analysis 2: Genetic Matching using standardized difference in means

Table 4.6 shows the treatment effect estimate when the standardized difference in means is used with GenMatch. The treatment effect is now 10% and 9% of standard deviation for samples 1 and 2, respectively. The difference here is that these estimates are statistically significant at the 5% level. This can be attributed to the different method used in the analysis. Unlike PSM, which weights observations (based on their propensity scores alone) to balance the propensity score density across treatment arms, GenMatch weights both the observations (matching) and the variables, to achieve balance in the propensity scores and the covariates. In the likely situation where balance in propensity score does not translate into balance in covariates, the results will be different. This is the advantage of the Genmatch approach. Instead of placing all the weights on the covariates (Mahalanobis matching) or all the weights on propensity scores (PSM), it finds the allocation of weights between the two approaches that improves balance optimally, given the data, the fitness function, and the measure of balance used. Its main disadvantage is that it can be tedious in terms of computation time.

Table 4.6 Treatment effect and Standard error for GenMatch (using standardized difference in means)

	Sample 1	Sample 2
Estimate	0.10641	0.090978
SE	0.046811	0.044105
T-stat	2.2733	2.0628
pval	0.023008	0.039135

4.5.3 Analysis 3: Genetic matching using KS distance as balance measure

This section and the next present the result of the analysis when distributional measures are used. In this section, I present the results when the KS statistic is used as the balance measure. It can be argued that the way the KS (and the entropy distance) will rank balance will be similar to the way standardized difference in means will rank balance for binary variables. Table 4.7 presents the results when the KS distance is used with GenMatch. The treatment effect is -2% and 10% for samples 1 and 2, respectively. The negative treatment effect in sample 1 is not significant, while the effect in sample 2 is significant at 5%. Results in sample 1 should be

interpreted with caution[33]. Results in sample 2 suggest that the KS distance is somewhat similar in performance to the standardized difference in means, in terms of effect size. However as we will show latter these results come from different optimal weights in the GenMatch algorithm. The crucial point is that different balance measures perform differently under the same method and using the same data. In other words, the balance measure does matter.

Table 4.7 Treatment effect and Standard error for GenMatch (using KS distance)

	Sample 1	Sample 2
Estimate	-0.0201	0.10599
SE	0.0433	0.0453
T-stat	-0.46426	2.3386
pval	0.64246	0.0193

4.5.4 Analysis 4: Genetic Matching using the entropy distance metric as a balance measure

This section presents the results when the entropy metric is used. First, I present the improvement in balance when the entropy measure is used. Note that, in the case of the entropy measure, this is defined for both discrete and continuous variables (See chapter 2). Table 4.8 below shows the entropy distance before and after Genmatch has been used to match the data. As one would expect, there is an improvement in all the covariates and the propensity scores after matching. This improvement occurs for dichotomous variables like marital status, and continuous variables like motivation. As noted earlier, it can be argued that the entropy metric will be similar to mean-based measures for binary variables. however, this is not the case for continuous variables. This distributional measure therefore accommodates diverse types of variables without ignoring any aspect of balance.

The greatest improvement in balance after matching and the greatest imbalance before matching occur in the continuous variables (i.e. motivation, and propensity scores). Specifically, the greatest imbalance before matching occurs in the propensity scores. The

[33] This warning is mainly because this result looks very different from the other results. A similar pattern was observed when I use the default setting that optimizes p-values of t-tests (for binary variables) and KS tests (for continuous variables). The treatment effect for the default setting is 4% and 14% of standard deviation for samples 1 and 2, respectively. However, this increases the amount of computation time considerably. It also means that the measures cannot be used directly to assess balance. Instead the p-values based on these measures are used.

treatment effect estimates from using the GenMatch for matching and the entropy metric as balance measures are presented in table 4.9 for both samples. The estimates are 11% & 16% of standard deviation in samples 1 and 2 respectively. We note that the difference in effect size in analyses 2, 3, and 4 can only be attributed to the balance measure.

Table 4.8 Entropy distance before and after matching

	A	B		C	D
variables in Sample1	**Before**	**After**	**variables in Sample 2**	**Before**	**After**
employed	0.097321	0.073682	**employed**	0.076828	0.061946
married	0.104859	0.073473	**married**	0.067702	0.061479
electricity	0.113014	0.067299	**electricity**	0.090005	0.057958
water	0.168985	0.052527	**water**	0.105215	0.055633
telephone	0.097727	0.036643	**hh_head_gen**	0.044779	0.037208
hh_head_gen	0.108057	0.032138	**cg_edu**	0.064754	0.036636
cg_edu	0.051905	0.01946	**cg_age**	0.027098	0.036518
cg_age	0.01171	0.004128	**coloured**	0.016045	0.02796
coloured	0.231005	0.002659	**Asian**	0.112151	0.020351
Asian	0.111662	0.000562	**Black**	0.184486	0.020081
Black	0.231005	0.000553	**telephone**	0.115864	0.004863
motivation	0.184061	0.000553	**motivation**	0.279157	0.003565
Motivation*employed	0.088273	0.000472	**Motivation*motivation**	0.217053	8.3E-05
Employed*hh_head_gen	0.098057	0.000377	**Motivation*employed**	0.112068	2.84E-05
cg_age*cg_edu	0.018981	3.78E-05	**cg_edu*married**	0.04	1.71E-05
Pscores	0.409029	0.00000	**hh_head_gen*married**	0.082493	1.63E-05
			Pscores	0.650965	0.00000

No direct comparison of the entropy measure with other measures of balance is carried out in this chapter. However, it is clear that using a different balance metric does lead to different results, as shown in tables 4.6, 4.7 and 4.9.

Table 4.9 Treatment effect and Standard error for GenMatch (using Entropy distance metric)

	Sample 1	Sample 2
Estimate	0.11389	0.16633
SE	0.045863	0.046443
T-stat	2.4833	3.5814
Pval	0.013016	0.000342

This is not surprising, since the balance measures capture different things. Our explanation for this is that the weight selected by GenMatch to optimize balance will depend on the balance metric used. The entropy measure and the KS statistic will lead to weights that attempt to balance the distribution of covariates, potentially accounting for problems of thin or no support. This is important for continuous variables like motivation and propensity score, and variables with more than two levels, in general. This contrasts with measures that seek to balance only the mean, or mean and variance. The latter will therefore ignore information that may improve balance in distribution and consequently influence the treatment effect.

Therefore, measures that ignore information about parts of the distribution may converge at weights that are suboptimal. The KS statistic also compares distributions. However, its performance in terms of effect size and the results in previous chapters suggest that it, too, might be converging at a sub-optimal point. This is because the way it compares distributions can lead it to ignore certain types of imbalance. Under the assumption that more balance leads to less bias, the entropy measure should be preferred. Furthermore, the path that the process will take to produce "offspring" solutions from "parent solutions" in GenMatch will vary with the balance measure used.

4.6 Influence of weights

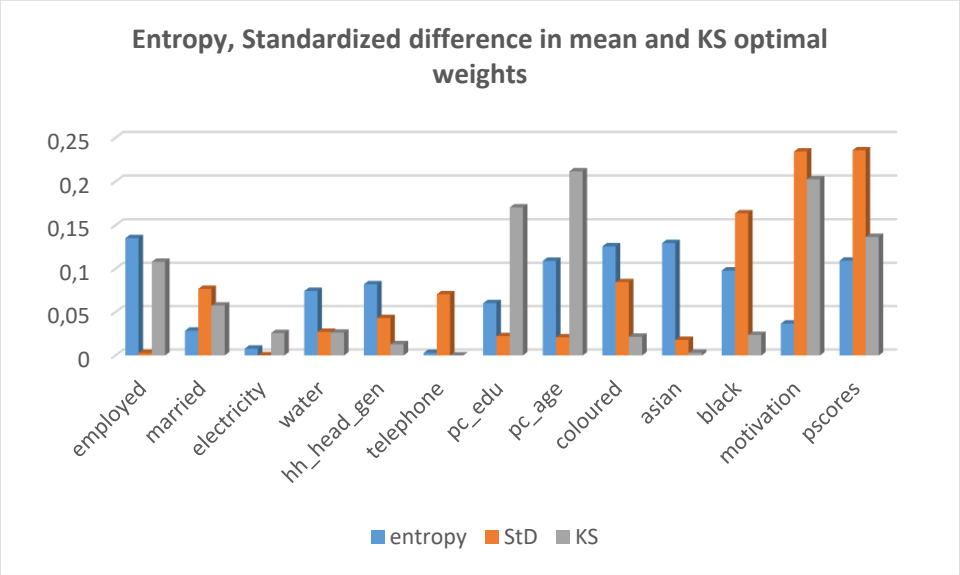
It was argued in the last chapter that when a large imbalance in the propensity score density, treatment effect will be influenced by the weights imposed by the econometric method. The analysis in the last section further suggests that, with the GenMatch approach, the balance measure used influences the weights and, therefore, the treatment effect estimate. We discuss further investigation of this in this section.

Figure 4.1 shows the relative weight[34] allocated to each variable under the balance measures for sample 2. It is clear that the weights are different. In theory, if all the covariates follow, for example a normal distribution, then the weights should be equivalent (across balance measures) since one can characterize normal distributions with their means and variances only. This will mean that the measures will capture similar information about balance. However, in practice, observing such a distribution is unlikely, so that the weights will differ,

[34] That is, weights are normalized to sum to 1.

and depend on the balance measure used. One glaring difference is that the standardized difference in means tends to allocate more weights to the continuous variables.

Figure 4.1: Distribution of weights



4.7 Precision of estimates

The analysis here and the one in Coetzee (2011) use standard error, which does not account for matching in drawing conclusions. While this makes it possible for me to compare my inference with the one in Coetzee (2011), it is less than ideal. Using the standard error of Abadie and Imbens (2004) that corrects for the fact that matching is used, the estimates from analysis 1, and analysis 2 sample 1 remain statistically insignificant, while other estimates become statistically insignificant (see tables 4.10, 4.11 and 4.12). What this suggests is that, although CSG tends to have a positive effect on the height-for-age z score, this effect is not measured with precision. This perhaps can be traced back to the data. The absence of application date data for many of the children in the control group means that censored regression must be relied on to calculate the motivation for these respondents. While this may help, it leaves a lot of room for improvement in terms of the number of control units used in the analysis (see table 4.4). This will have a direct impact on the standard error of the estimates. More observations with application data in the control sample may improve the sample size in control sample 1, and consequently improve the precision of the measurement.

Table 4.10 Treatment effect and Abadie and Imbens Standard error for GenMatch (using standardized difference in means)		
	Sample 1	Sample 2
Estimate	0.10641	0.090978
AI SE	0.31182	0.3153
T-stat	0.34127	0.2885

Table 4.11 Treatment effect estimate and Abadie and Imbens Standard errors (using KS statistic)		
	Sample 1	Sample 2
Estimate	-0.0201	0.10599
AI SE	0.0434	0.25797
T-stat	-0.46426	0.41087

Table 4.12 Treatment effect and Abadie and Imbens Standard error for GenMatch (using Entropy distance metric)		
	Sample 1	Sample 2
Estimate	0.11389	0.16633
AI SE	0.27575	0.25464
T-stat	0.41303	0.6532

4.8 Conclusion

This chapter uses the NIDS data to illustrate one possible use of the entropy metric as a balance measure. Following Coetzee (2011), it draws treatment and control samples from the NIDS wave 1 data for 2008. However, several changes are made to the analysis. These changes lead to the conclusion that the CGS has a positive impact on the height-for-age z scores of children whose caregivers received the grant for at least 34% of the childrens' lives, compared to children who qualify but are not part of the programme. The main changes include, firstly, using a different method to calculate caregiver motivation, to make sure that all the data is employed. Secondly, I use GenMatch for matching, which is expected to achieve a more reliable level of balance than the popular PSM approach (Sekhon, 2011). Finally, I redefine the treatment group to mitigate the effect of diluting the treatment effect with treated observations which have a low dosage of the CSG.

The treatment effect estimate reported here and in Coetzee (2011) suggests that the CSG has a positive impact on height-for-age scores. However, the estimates are not significant when

standard error takes the matching process into account. I argue that this can be improved by getting information on application data for eligible non-beneficiaries.

Finally, the analyses reported in this chapter shows that the balance measure used to calculate the treatment effect does matter for the result under GenMatch. Using a measure that considers covariate distribution rather than a few moments leads to stronger effect estimates concerning the size of the effect.

5 Concluding remarks

This thesis examines the balancing condition in evaluation studies and introduces a new measure that can be used to better quantify balance. The empirical work shows that the measure used to assess balance does matter. I argue that assessing balance using a few moments, which is a popular method in the literature, leaves room for improvement. This is because measures that use only a few moments to assess balance ignore information from other parts of the covariate distribution. It can be argued that balance at the first and second moments should be sufficient to recover consistent estimates. However, results shown in this thesis suggest that considering the entire distribution when assessing balance is important. This is true if one is interested in an analysis that can replicate experimental results, or wants to have an idea of how close a given analysis is to replicating experimental results, in terms of bias and robustness.

Chapter 2 of our study provides analyses which give a reason why distributional measures should be preferred. In this chapter, it is shown that it is possible to redistribute mass on the densities being compared so that both the mean and the KS statistic are blind to the effect of the redistribution. Specifically, the example in chapter 2 shows that one can configure redistribution of masses on common support so that the thin/no support problem is increased within the support, without upsetting the balance assessments of the mean and the KS statistic.

The results of the performance of the KS statistic and the entropy measure should not be a surprise, if one considers the way that "difference" in distribution is quantified under both measures. The entropy measure considers the entire support, giving equal weighting to differences wherever they occur on the support. On the other hand, the KS statistic is based on the maximum distance between the cumulative distribution functions of the densities being compared. This suggests that the region of support with the greatest difference will drive results under the KS statistic. In addition to this, the literature on the KS statistic across fields suggests that this measure is more sensitive to deviations at the centre of the distribution, at the expense of deviations at the tails (Kole *et. al,* 2007; Parizzi & Brcic, 2011; Kaplan & Goldman, 2015). This is clearly a problem when assessing imbalance that may manifest as a thin/no support problem, because these forms of imbalance are more likely to occur at the tails.

The proposed entropy measure provides better information about balance than the alternatives (standardized difference in means and KS statistic). One should therefore expect it to be better correlated with the size of the bias. The analysis discussed in chapter 3shows that, when balance is assessed with propensity score density, one can better discriminate between the levels of balance achieved by plausible control samples, using the entropy measure. This discrimination is based on the correlation between balance measures and the size of the bias (relative to the experimental benchmark estimate). The entropy measure also performs better in predicting the robustness of effect estimates across econometric methods. In the case of robustness, one can think of imbalance in propensity score density as a gap that will be inadvertently filled by the econometric approach or matching method used to estimate treatment effect. If there is a small gap or no gap to be filled (adequate balance), then the econometric approach used to estimate treatment effect will have little influence on the results. Otherwise, each econometric approach will fill the gap differently--subject to its assumptions, so that treatment effect will vary across methods. Any balance measure that does not quantify the difference between distributions accurately will fail to capture the correlation between balance, bias, and robustness of effect estimates. Therefore, when a number of plausible control samples are available, using the entropy measure to select the control sample that is closest to the treatment sample, as measured by the entropic distance, can be a useful way to mitigate bias.

The analysis in Chapter 4 shows that, when entropy is used as a balance measure, a stronger effect size (16% of standard deviation) is found, compared to when the standardized difference in means is used (10% of standard deviation in sample 2). This is important, because even though both estimates are statistically significant, there is a slight difference in the strength of the conclusions. Tables 4.6A and 4.8A show that (ignoring the matching method) the standard error of the treatment effect estimates under the two balance measures are similar, so that the t-statistic is larger under the entropy measure (2.06 vs 3.58). For a programme like the Child Support Grant, this is important, because certain research based on NIDS data suggests that treatment effect of the grant on a wide range of outcome variables that relate to child welfare is not significant, in the binary case (Coetzee, 2013 & 2011). Our results show a significant effect for the height-for-age z score when GenMatch is

used. This highlights the fact that the previous results (which are based on PSM) may not be conclusive.

There are some caveats to the use of the proposed measure to assess balance. In this thesis, the entropy measure is used in situations where interest is in balance in one sample compared to another (comparable) sample. In other words, the measure is used to assess relative balance, rather than balance in absolute terms. For example, in the case of standardized difference in means, values above 0.2 are generally regarded as indicative of too much imbalance. Based on the analyses in this thesis, it is difficult to imagine a figure beyond which the entropy metric would be indicative of a situation where imbalance is too high (either in the univariate case or when propensity score density is used). The results discussed in chapter 3 suggest that an entropic distance between propensity score densities that is greater than 0.1 may be indicative of a worrisome level of imbalance. However, more research is needed to establish this. I can cautiously say that, when balance is measured on the propensity score density, and the specification used to estimate the propensity scores satisfies the DW balancing condition, an entropic distance of 0.1 between propensity score densities is a reasonable figure to keep in mind. Note that this does not mean that, when the entropic distance is greater than 0.1, the treatment effect cannot be estimated on the sample. What this means is that the treatment effect may depend on the method used in its estimation, in terms of bias and robustness.

Furthermore, like every balance measure, the entropic distance relies heavily on the selection on observables assumption. Its ability to capture balance in a way that is most informative about bias relies on the assumption that every relevant variable is included in estimating the propensity scores. It can only measure balance in observed attributes, therefore, if a relevant attribute is missing in the propensity score specification, the result may be inconsistent in providing informative about bias. This is more important when we are thinking about balance in absolute terms. The research covered in chapter 3, for example is concerned with relative balance in each sample. One can therefore safely assume that the effect of an unobserved variable is constant across samples, so that the comparison will still be consistent. The same argument can be made about the analyses discussed in chapter 4, which compares balance in one weighted sample to balance in another weighted sample. Across these comparisons

(all things being equal), the effect of an unobserved variable is constant, so comparison can still be made.

5.1 Recommendations and Further Research

As shown by its implementation discussed in chapters 3 and 4, the entropic measure is particularly useful when one has different comparable estimation samples to estimate the same treatment effect (fixed treatment sample). The general idea is that there is an ideal level of balance, but there are several ways to select the control sample to be used in estimation, and there is no information on which method achieves the ideal level of balance. My research in chapter 3 is based on a theory concerning how the control sample should be selected. However, this does not specify how many years of recent employment history are necessary to balance our treatment sample. This is an empirical question, as there is no guarantee that the same amount of recent employment history will work under different data sets. Including as many observations as possible may increase efficiency. However, when it comes to the trade-off between bias and efficiency, there is no point in having a precise estimate of a quantity that is wrong (Rubin, 2006). One can think of the entropy metric as a way of screening out observations that can be considered as "outliers", in this case, in the PSID and CPS datasets. This is because sample 2 is a subsample of sample 1, and sample 3, in turn, is a subsample of sample 2. The entropy measure in this example shows that the assumptions that lead to the selection of samples 1 and 2 are not sufficient to achieve balance in distribution, with respect to the treatment sample under consideration. The DW algorithm helps in getting rid of some "outliers" in each sample. However, the result shows that there are still systematic differences in the pre-screened samples that can be traced back to the level of balance achieved. The set-up for the analysis covered in chapter 4 is different. However, the logic of differentiating between different samples that can be used to estimate the same treatment effect is similar to the idea covered in chapter 3. One can think of the weighted samples in all stages of the GenMatch optimization as different plausible samples to estimate the effect of the CSG. In both cases, balance in one sample, compared to another, is what is important.

There are other situations where a similar set-up may exist. For example, in a regression discontinuity approach, all that is known is that treatment and control observations should be selected around the point of discontinuity. The treatment observations can be fixed and

an optimal distance from the point of discontinuity chosen for selecting control observations without compromising balance in observables. There is often a sample size problem which temps one to include observations further away from the cut-off in the sample, to improve efficiency. The entropic measure can be used to gauge when balance is being compromised.

Re-randomization checks is another situation that fits this description. Randomized experiments may be ideal in estimating causal inference. However, chance imbalance may occur after randomization between covariate distributions across treatment arms. The probability of observing a large imbalance across treatment arms falls with sample size (Bruhn & Mckenzie, 2009). One approach to fix this problem is through re-randomization. Unbalanced randomization can be discarded, followed by re-randomization, provided a precise definition of imbalance has been specified before an experiment (Morgan and Rubin, 2012).. This process can be continued until a randomization that yields balance according to the pre-specified definition of imbalance is achieved (see Morgan & Rubin (2012) and Bruhn & Mckenzie (2009) for a more detailed description). One can relate this practise of re-randomization until a pre-specified level of balance is reached to what is done under GenMatch. The latter changes the sample (after treatment) by weighting covariates to achieve the optimum level of balance possible, as shown in chapter 4.The former changes the sample (before treatment) by re-randomizing until a pre-specified level of balance is achieved. In both cases, balance in one potential estimation sample is compared to balance in another potential estimation sample. Note that, in both cases, the sample size is fixed.

More specific examples can be given on how the entropy measure can be used for re-randomization. Consider two methods that involve multiple randomization described by Bruhn & Mckenzie (2009). The goal in both cases is to increase the likelihood of balance on observed characteristics. The first involves taking a random draw of assignment to treatment, examining the difference in means for the covariates, and then re-randomizing if the difference is too large. The second method takes many draws of treatment assignment, and then selects the one that exhibits the greatest level of balance. The pre-specified condition can be a function of the t-statistic or p-value of mean difference of t-tests. Instead of using mean based statistics, one can specify the balancing condition as a function of the entropy distance, or the p-value of the entropy distance, between covariate distributions. This is, of course, a stricter condition than the mean-based approach, but it has the advantage of providing more information about balance. If a weaker pre-specified balancing condition is

used with the entropy measure, the researcher is at least aware of the extent of imbalance in the sample. This should be preferred to a situation where the balance measure may be blind to some forms of imbalance. As mentioned earlier, I highlight these as points that can be considered in future research.

Bibliography

Abadie, A., & Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica, 76*(6), 1537-1557.

Abadie, A., & Imbens, G. W. (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business \& Economic Statistics, 29*(1).

Aguero, J., Carter, M., & Woolard, I. (2006). The impact of unconditional cash transfers on nutrition: The South African Child Support Grant.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics*, 47-57.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine, 28*(25), 3083-3107.

Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics, 50*(4), 495-510.

Becker, S. O., Ichino, A., et al.. (2002). Estimation of average treatment effects based on propensity scores. *The stata journal, 2*(4), 358-377.

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H., Boer, A., & Klungel, O. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and drug safety, 20*(11), 1115-1129.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhy: The Indian Journal of Statistics*, 401-406.

Blundell, R., & Dias, M. C. (2002). Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal, 1*(2), 91-115.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics, 1*(4), 200-232.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys, 22*(1), 31-72.

Case, A., Hosegood, V., & Lund, F. (2005). The reach and impact of Child Support Grants: evidence from KwaZulu-Natal. *Development Southern Africa, 22*(4), 467-482.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Cha, S.-H. (2007). Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions. *International Journal of Mathematical Models and Methods in Applied Sciences.*

Coetzee, M. (2011). Finding the Benefits: Evaluating the Impact of the South African Child Support Grant. *Economic Research Southern Africa*, (p. 7).

Coetzee, M. (2013). Finding the benefits: Estimating the impact of the South African child support grant. *South African Journal of Economics, 81*(3), 427-450.

Dehejia, R. (2005). Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics, 125*(1), 355-364.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association, 94*(448), 1053-1062.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics, 84*(1), 151-161.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945.

Diamond, A., & Sekhon, J. S. (2013). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics, 95*(3), 932-945.

Duflo, E. (2003). Grandmothers and Granddaughters: Old-Age Pensions and Intrahousehold Allocation in South Africa. *The World Bank Economic Review, 17*(1), 1-25.

Ebrahimi, N., Maasoumi, E., & Soofi, E. S. (1999). Ordering univariate distributions by entropy and variance. *Journal of Econometrics, 90*(2), 317-336.

Gelman, A., & Hill, J. (2007). Causal inference using more advanced models. *Data analysis using regression and multilevel/hierarchical models (1st ed), Cambridge University Press, New York*, 215-226.

Goldman, M., & Kaplan, D. M. (2016). *Evenly sensitive KS-type inference on distributions.* Tech. rep., Working paper, available at http://faculty. missouri. edu/\~ kaplandm .

Granger, C., Maasoumi, E., & Racine, J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis, 25*(5), 649-669.

Hainmueller, J. (2012). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis, 20*, 25-46.

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching As an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme. *The Review of economic studies, 64*(4), 605-654.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*(5), 1017-1098.

Hill, J., Su, Y.-S., & others. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics, 7*(3), 1386-1420.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*, 199-236.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis* (Vol. 3). Wiley New York.

Iacus, S. M., King, G., Porro, G., & Katz, J. N. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, 1-24.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society), 171*(2), 481-502.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Kole, E., Koedijk, K., & Verbeek, M. (2007). Selecting copulas for risk management. *Journal of Banking & Finance, 31*(8), 2405-2423.

Kvam, P. H., & Vidakovic, B. (2007). *Nonparametric statistics with applications to science and engineering* (Vol. 653). John Wiley & Sons.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.

Lechner, M., & Strittmatter, A. (2014). *Practical Procedures to Deal with Common Support Problems in Matching Estimation.* Tech. rep., University of St. Gallen, School of Economics and Political Science.

Lee, W.-S. (2013). Propensity score matching and variations on the balancing test. *Empirical economics, 44*(1), 47-80.

Leuven, E., Sianesi, B., & others. (2015). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. *Statistical Software Components*.

Li, Q., Maasoumi, E., & Racine, J. S. (2009). A nonparametric test for equality of distributions with mixed categorical and continuous data. *Journal of Econometrics, 148*(2), 186-200.

Maasoumi, E., & Racine, J. (2002). Entropy and predictability of stock market returns. *Journal of Econometrics, 107*(1), 291-312.

Maasoumi, E., & Racine, J. S. (2008). A robust entropy-based test of asymmetry for discrete and continuous processes. *Econometric Reviews, 28*(1-3), 246-261.

Maasoumi, E., Wang, L., & others. (2012). *The gender earnings gap: Measurement and analysis.* Tech. rep., Working paper.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta), 2*, 49-55.

Matusita, K. (1955). Decision rules, based on the distance, for problems of fit, two samples, and estimation. *The Annals of Mathematical Statistics*, 631-640.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 1263-1282.

Parizzi, A., & Brcic, R. (2011). Adaptive InSAR stack multilooking exploiting amplitude statistics: A comparison between different techniques and practical results. *IEEE Geoscience and Remote Sensing Letters, 8*(3), 441-445.

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24-43.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics*, 109-120.

Rubin, D. B. (2006). *Matched sampling for causal effects.* Cambridge University Press.

Schweizer, B., Sklar, A., et al. (1960). Statistical metric spaces. *Pacific J. Math, 10*(1), 313-334.

Sekhon, J. S. (2006). Alternative balance metrics for bias reduction in matching methods for causal inference. *Work. Pap., Dep. Polit. Sci., Univ. Calif. Berkeley*.

Sekhon, J. S., & Diamond, A. (2005). Genetic matching for estimating causal effects. *Unpublished Manuscript. Presented at the Annual Meeting of the Political Methodology, Tallahassee, FL.*

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in medicine, 13*(17), 1715-1726.

Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of experimental criminology, 9*(2), 129-144.

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell Labs Technical Journal, 28*(4), 656-715.

Skaug, H., & Tj{\o}stheim, D. (1996). Measures of distance between densities with application to testing for serial independence. *Time Series Analysis in Memory of EJ Hannan, New York: Springer*, 363-377.

Smith, J. (2000). *A critical survey of empirical methods for evaluating active labor market policies.* Tech. rep., Research Report, Department of Economics, University of Western Ontario.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of econometrics, 125*(1), 305-353.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score--based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology, 66*(8), S84--S90.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference, 49*(1), 137-162.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology, 42*(1), 314-347.

Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence.*Review of economics and statistics, 86*(1), 91-107.

Appendix

A1 Heckman .et al. bias decomposition

$$B_{Q1} = \int_{S_1} E(Y_0|W, D = 1)f(W|D = 1)dW - \int_{S_0} E(Y_0|W, D = 0)f(W|D = 0\ )dW$$

= (i) - (ii)

Here I am splitting the support

(i) =

$$\int_{S_1 \setminus S_{10}} E(Y_0|W, D = 1)f(W|D = 1)dW\ (call\ this\ H)$$

+

$$\int_{S_{10}} E(Y_0|W, D = 1)f(W|D = 1)dW\ (call\ this\ I)$$

(ii) =

$$\int_{S_1 \setminus S_{10}} E(Y_0|W, D = 0)f(W|D = 0)dW\ (call\ this\ J)$$

+

$$\int_{S_{10}} E(Y_0|W, D = 0)f(W|D = 0)dW\ (call\ this\ K)$$

Now we can add and subtract

$$\int_{S_{10}} E(Y_0|W, D = 0)f(W|D = 1)dW\ (call\ this\ L)$$

So that $B_{Q1} = a + b + c$

let

$$b = H - J, c = -K + L \; and \; a = I - L$$

So that $a + b + c = H - J - K + L + I - L$

L Cancels out

$$a + b + c = H - J - K + I$$

$$a + b + c = (i) - (ii)$$

Where (b) is the component due to differing supports of W, (c) is the component due to differing distribution of W over the same support in the two populations and (a) is the component due to differences in outcomes that are present, even after controlling for observables. We can then write

$$b = \int_{S_1 \setminus S_{10}} E(Y_0|W, D = 1) f(W|D = 1) dW - \int_{S_1 \setminus S_{10}} E(Y_0|W, D = 0) f(W|D = 0) dW$$

$$c = -\int_{S_{10}} E(Y_0|W, D = 0) f(W|D = 0) dW + \int_{S_{10}} E(Y_0|W, D = 0) f(W|D = 1) dW$$

$$a = \int_{S_{10}} E(Y_0|W, D = 1) f(W|D = 1) dW - \int_{S_{10}} E(Y_0|W, D = 0) f(W|D = 1) dW$$

Which can then be written as

$$b = \int_{S_1 \setminus S_{10}} E(Y_0|W, D = 1) f(W|D = 1) dW - \int_{S_1 \setminus S_{10}} E(Y_0|W, D = 0) f(W|D = 0) dW$$

$$c = \int_{S_{10}} E(Y_0|W, D = 0) \{f(W|D = 1) - f(W|D = 0)\} dW$$

$$a = \int_{S_{10}} \{E(Y_0|W, D = 1) - E(Y_0|W, D = 0)\} f(W|D = 1) dW$$

A2 Data that produced the results in section 2.5

Appendix Table 1: Probability mass function in scenario A, B and C

support		**Scenario A**		**Scenario B**		**Scenario C**	
	x	f_0	f_1	f_0	f_1	f_0	f_1
L	0.00	0.050	0.122	0.053	0.020	0.059	0.020
M	2.13	0.167	0.196	0.010	0.240	0.00	0.240
N	5.48	0.685	0.273	0.834	0.327	0.838	0.327
O	6.20	0.048	0.210	0.053	0.161	0.053	0.161
P	8.00	0.050	0.199	0.050	0.252	0.050	0.252
		1.00	1.00	1.00	1.00	1.00	1.00

Appendix Table 2: Cumulative distribution function

support		**Scenario A**		**Scenario B**		**Scenario C**	
	x	F_0	F_1	F_0	F_1	F_0	F_1
L	0.00	0.05	0.12	0.05	0.02	0.06	0.02
M	2.13	0.22	0.32	0.06	0.26	0.06	0.26
N	5.48	0.90	0.59	0.90	0.59	0.90	0.59
O	6.20	0.95	0.80	0.95	0.75	0.95	0.75
P	8.00	1.00	1.00	1.00	1.00	1.00	1.00

Appendix Table 3: Entropy distance & KS statistic

Scenarios		**A**	**B**	**C**	**A**	**B**	**C**
support	x	$(f_1 - f_2)^2$			**|F1-F2|**		
L	0.00	0.02	0.01	0.01	0.07	0.03	0.04
M	2.13	0.00	0.15	0.24	0.10	0.20	0.20
N	5.48	0.09	0.12	0.12	*0.31*	*0.31*	*0.31*
O	6.20	0.06	0.03	0.03	0.15	0.20	0.20
P	8.00	0.05	0.08	0.08	0.00	0.00	0.00
S_ρ & KS		**0.22**	**0.39**	**0.48**	**0.31**	**0.31**	**0.31**

Appendix Table 4: Mean

		Scenario A		Scenario B		Scenario C	
support	x	$x * f_0$	$x * f_1$	$x * f_0$	$x * f_1$	$x * f_0$	$x * f_1$
L	0.00	0.00	0.00	0.00	0.00	0.00	0.00
M	2.13	0.36	0.42	0.02	0.51	0.00	0.51
N	5.48	3.75	1.49	4.57	1.79	4.59	1.79
O	6.20	0.30	1.30	0.33	1.00	0.33	1.00
P	8.00	0.40	1.59	0.40	2.02	0.40	2.02
	mean	4.81	4.81	5.32	5.32	5.32	5.32

B1 Propensity score specifications

Appendix Table 5 Propensity Score specification for PSID and CPS Samples

TREATTMENT VS CONTROL	
	LALONDE SAMPLE
Lalonde PSID1	age education married nodegree black hispanic re75 married*re75 hispanic*re75 age2 nodegree*black
Lalonde PSID2	age education married nodegree black hispanic re75 married*re75 hispanic*re75 age2
lalonde PSID3	age education married nodegree black hispanic re75 age2 edu2
Lalonde CPS1	age education married nodegree black hispanic re75 age2 edu2 re752
LalondeCPS2	age age2 education married nodegree black hispanic re75 nodegree*education Hispanic*re75
LalondeCPS3	age age2 education married nodegree black hispanic re75 nodegree*education
	DW SAMPLE
DW PSID1	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black*z74
DW PSID2	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 nodegree_re75 edu*re74 married*re75
DW PSID3	age education married nodegree black hispanic re75 re74 age2 edu2 nodegree_re74
DW CPS1	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black_z74
DWCPS2	age education married nodegree black hispanic re74 re75 age2 edu2 re742 re752 black_z74
DWCPS3	age education married nodegree black hispanic re74 re75 age2 edu2 re752

Age: Age in years
education: years of education
black Hispanic: dummy variables for the race of the respondent
married: dummy variable for marital status.
re75: real income in 1975
Note: age2, edu2: square of age and square of education and "*" denote interaction of terms e.g. hispanic*re75 is the interaction of race with income in 1975.

C1 Number of observations *in table 4.3*

Appendix table 6: Number of observations *in table 4.3*							
	Low	**medium**	**High**	**All treatment**	**All Control**	**Eligible Control**	**Ineligible Control**
Caregiver Characteristics							
Motivation	714	763	1015	2492	107	107	0
delay_yr	714	763	1015	2492	107	107	0
Employed	692	727	982	2401	1199	1199	597
Married	714	763	1012	2489	1280	1280	657
Education	714	762	1014	2490	1279	1279	655
Age	714	763	1015	2492	1285	1285	655
Child Characteristics							
Age	714	763	1015	2492	1286	1286	657
Gender	714	763	1015	2492	1286	1286	657
Black	714	763	1015	2492	1286	1286	657
Coloured	714	763	1015	2492	1286	1286	657
Asian	714	763	1015	2492	1286	1286	657
White	714	763	1015	2492	1286	1286	657
HH Characteristics							
Electricity	714	763	1015	2492	1286	1286	657
Water	714	763	1015	2492	1286	1286	657
Telephone	714	761	1011	2486	1286	1286	656
Toilet	710	761	1014	2485	1284	1284	652
HH head Gender	672	705	968	2345	1212	1212	633