An Evaluation of Balance Metrics in Propensity Score Analyses

By

Patrick O’Keefe

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Psychology
May 8th, 2020
Nasvhille, Tennessee

Approved:
Dr. Joseph Lee Rodgers
Dr. Kristopher J. Preacher
Dr. Andrew J. Tomarken
Dr. David Cole

Table of Contents

Page

List of Figures

Figure Page

List of Tables

Table Page

Chapter 1

Introduction

Propensity score analysis is a method for creating matched groups of treated and control subjects from otherwise unmatched samples (e.g., Rosenbaum & Rubin, 1983). Frequently researchers will encounter situations where it would be impossible or unethical to rely on more traditional experimental methods (e.g., random assignment). Using propensity score methods allows researchers to obtain some of the same benefits from a quasi-experiment as from a randomized study or a randomized control trial. For example, if a researcher wished to compare the effects of school type (public or private) on standardized test scores, propensity score analysis would provide a means to do so. It is nearly a truism that students at public and private schools are different on a number of different demographic variables, yet there are likely students in public schools that are quite similar to students in private schools and vice versa. A naïve approach would be to exactly match students on confounding characteristics, however this approach could result in very few matches. The key statistical insight of propensity score analysis is that if "treatment" (here public vs. private school attendance) can be predicted by some propensity for treatment, the propensity score, then if subjects are matched on that propensity score the matched groups ought to have similar multivariate distributions on the confounds. This result only occurs if the propensity score is correctly calculated. The benefit of this method of matching, above and beyond other forms of matching, is that individuals only need to be matched on a single measure, the propensity score, in order to be considered comparable, which makes finding matches much easier.

The basic procedure in a propensity score analysis is to create some model of the propensity score. The researcher calculates the propensity score and then matches subjects on this propensity score. After matching is done unmatched subjects are discarded. The most common way to obtain a propensity score is to create a logistic regression model predicting treatment status (e.g., enrollment in public vs. private school) using potential confounds as predictors. Confounds are those variables that might predict both treatment status as well as the outcome of treatment (e.g., standardized test scores). The model used to predict treatment status is the propensity score model, and is typically a logistic regression model, but researchers are not necessarily limited to logistic regression, and may use any statistically valid classification technique as a propensity score model.

Propensity score analysis may be less familiar to researchers than other statistical methods of managing confounds, like regression or analysis of covariance, yet it serves a similar purpose. Each method could be used in a statistical design to account for potentially confounding variables in order to assess the effect of a treatment above and beyond those confounds. Propensity score analysis is arguably a less flexible method, in that it is largely useful for managing confounding, whereas regression and other linear models can be used to simultaneously assess multiple effects of interest. Despite this relatively narrow focus, propensity score analysis has some practical benefits that set it apart from more traditional approaches such as regression analysis. Perhaps the most underappreciated is that propensity score analysis completely excludes the ultimate outcome from the initial stages of the analysis (e.g., Ho, Imai, King, & Stuart, 2007). The propensity score process is siloed from the actual analysis of interest, and is specific to every sample. This means a researcher generally ought to be unconcerned about the "population" propensity score model, and should be more concerned about what differences exist between treated and control individuals in their data sample. The propensity score model is expected to vary from sample to sample, and that variation means that the propensity score needs to change from sample to sample (e.g., Dehjia, 2005). In order to account for this natural variability, researchers will engage in a model building process where the initial model is fit, matches are created, balance is assessed, and adjustments to the model are made to create better balance. This iterative process could be concerning except that the ultimate analysis of interest is not the propensity score model, but is the outcome model. As the outcome of interest is excluded from this model building exercise, any capitalization on chance should have minimal effect on the results of the outcome model. This kind of model building exercise cannot be undertaken in regression analysis without creating issues of multiplicity and *p*-hacking.

There are three basic steps to propensity score analysis. In the first step, calculating, a propensity score is created for each individual. This propensity score is a measure of the probability (or odds or likelihood) of a given individual having received treatment. Typically this propensity score is calculated from a logistic regression model predicting the log-odds of treatment using the measured confounds. In the second step, matching, an individual from the treated group is matched with an individual from the control group. Once all the matches have been created the third step, assessing balance, assesses balance between the matched treated and control groups on the covariates.

Rosenbaum and Rubin (1983) showed that if a person's expected outcome under the hypothetical condition that they were treated (or their expected outcome under the hypothetical condition that they did not receive treatment) is independent of their actual treatment status given a set of covariates **X**, then, if treatment is conditioned on a propensity score created using **X,** the expected outcomes will also be independent of treatment. This discussion regarding expected outcomes of treatment being independent of treatment can be somewhat confusing. Consider students from private and public schools. A researcher is interested in comparing their SAT scores. The treatment here is school type. We might expect a number of socioeconomic variables to differ between the children in these two schools, features which are not part of the school experience itself but rather due to selection effects. If we condition on all the relevant socioeconomic variables (e.g., family income, parental education; the **X** variables in this example), we would expect that the effect of moving a student from public to private schools would be the same for all students **regardless of their actual school status** and vice versa. So consider Jack and Jim. Jack is in public school and Jim is in private school. If we have conditioned on the relevant confounding variables we would expect the effect of switching from public to private school to be the same for both Jack and Jim, despite the fact that their reality is that they are in different schools. Propensity score analysis allows us to take what might be a very large **X** vector of confounds and collapse it into a single propensity score.

Calculation

Calculating a propensity score is relatively straightforward. Most research uses logistic regression and predicts treatment from the set of possible confounds. Rosenbaum and Rubin (1984) in one of the earliest demonstrations of propensity score analysis used a logit model. However, alternatives to simple logistic regression exist and the literature supports their use (e.g., Harder, Stuart, & Anthony, 2010). As an example of how the propensity score is calculated, consider the example of public versus private schools. If family income is different between public and private school children, then income will be a predictor of school type.

Interestingly this will also be true of variables that are not confounds per se, but simply unbalanced due to chance, or variables that are predictive of treatment but not predictive of the outcome. So height could be different between public and private school children, and because of this imbalance it would also be a predictor of school status, but it is difficult to think of how height could predict SAT scores. This illustration shows that not all variables that predict treatment status are confounds, only variables that predict both treatment status and the outcome of interest are confounds. Furthermore, some variables that are highly predictive of the outcome may be balanced between treated and control units and thus will not be a predictor of the propensity score (however, this situation is entirely reasonable as an already balanced predictor will not influence our estimate of the effect of treatment). Again, this variable is not a confound because, while it predicts the outcome of interest, it is unrelated to treatment status.

Matching

Once a suitable model for the propensity score has been created, the propensity scores are calculated for each individual. Typically, the next step is to match individuals on the propensity score. Researchers are not limited strictly to matching, however I believe that it is conceptually the easiest method to use and also the most theoretically sound. There are many ways to create matches in propensity score analysis (e.g., Austin, 2014) but the simplest is 1:1 nearest neighbor matching. This is the procedure that will be used in this project. In this procedure an individual from the treatment group is selected at random and paired with the individual in the control group with the closest match on the propensity score. This pair is added to the pool of matched subjects, and a new individual is randomly selected from the pool of treated individuals to be matched to a control individual. This process is repeated until the pool of treated individuals is exhausted. Matching procedures can be extended. One extension is m:n matching, where some number of treated subjects that are close on the propensity score are matched to some (not necessarily equal) number of control subjects who are also similar on the propensity score. Another extension is the use of a caliper. A caliper is defined as some distance on the propensity score (e.g., ±0.1 sd on the propensity score), then if no control subject can be found that is within that distance of a selected treated unit the treated unit is discarded. The use of a caliper can prevent bad matches from being created, however it can also limit the number of viable matches. In addition to matching techniques, the propensity score can be used to stratify subjects (stratification), and the outcome model fit to each strata. The propensity score can be used to weight the data, to make the treated and control groups more comparable (propensity score weighting). Finally, the propensity score itself can be used as a covariate in an ANCOVA type model (adjustment via the propensity score). All these methods are accepted practice within the propensity score literature, however for the present paper I focus on 1:1 nearest neighbor matching as I believe it is often used, conceptually the simplest, and provides adequate results.

Balancing

After matching is completed, the “balance” of the matched treated and control groups is tested. Typically, balancing takes the form of calculating the Cohen’s *d* of treatment for each confounding variable (e.g., Austin, 2008). A threshold of 0.1 appears to be used frequently, however this does not appear to be based on methodological research, and seems to be simply common practice. If balance is found across all variables, then a t-test on the outcome comparing treated to controls can proceed. If balance is not found, then the propensity score model might be changed, with other variables or higher order polynomial terms for existing variables added.

Less frequently, the matching scheme may be changed to attempt to achieve better balance. This process of model fitting, matching, and assessing balance, is repeated until adequate balance is found.

Motivation for the current paper

The potential for imbalance brings us to the primary motivations of this paper. First, does picking the model with the best balance consistently produce the best result? Second, are some balance tests better than others in this regard? Work by Diamond and Sekhon (2012) has resulted in an R package, Genmatch, that uses an iterative balancing method to achieve optimal match. However, this software method still leaves researchers to debate which balance metric they ought to optimize. Furthermore, in their work Diamond and Sekhon focus on univariate balance statistics, which may be inadequate. The inadequacy of univariate measures of balance, and the lack of research on multivariate measures of balance, is the final motivation for this paper.

Current practice

Austin (2009) made several recommendations regarding assessing balance. The main suggestions were that researchers use Cohen’s *d* to assess mean balance and balance on interaction terms, to also assess the balance of variance using variance ratios, and to compare Q-Q plots. In actual practice, when balance tests are reported, it appears that Cohen’s *d* is most often used to assess balance. In the field of propensity score analysis there is an avoidance of statistical significance tests for evaluating balance, although this avoidance is not universal; some researchers argue that statistical significance tests are unwise in this context (e.g., Austin

2008; Ho et al., 2007). The primary argument against using statistical tests rests on two issues in propensity score analysis specifically. First, there would be a perverse incentive for researchers to use smaller samples, and even to constrain the matching procedure to artificially limit the sample size. This research behavior would occur because, all else being equal, as sample size increases so does the power to reject the null hypothesis of no difference, therefore as sample size increased it would become more difficult to find a set of matched controls that was not statistically different from the treated sample. In addition, researchers may find small differences to be tolerable. The second issue, is that the use of a statistical test presupposes some population distribution to compare to, yet the propensity score method is a sample focused procedure (i.e., the goal is to create a balanced sample regardless of the true population). Not all researchers agree with this logic (e.g., Hansen, 2008), however it appears to be the dominant stance among propensity score practitioners. Cohen's *d* solves both problems. It gives researchers a statistically sound method to compare the two groups, and also allows for some small differences that do not penalize a researcher for having a large sample (in fact, one can show that in smaller samples Cohen's *d* would have greater variability and so smaller samples are ***more*** likely to result in unacceptable balance when balance exists in the population).

This issue bears at least some resemblance to measures of fit in structural equation modeling (SEM). In SEM it is frequently true that the model is acceptable, but is incorrect in the population in some small way. With large samples in SEM the χ^2 deviance statistic will nearly always be statistically significant, even if the model is "good enough". As a result, in SEM analyses unacceptably small samples appear to fit well (even if the model is wildly inaccurate) whereas SEM analyses with large samples may appear to fit poorly (even if the model is very close to the true population model). The SEM literature has compensated by deriving a number of fit indices that are designed to account for the penalizing influence of large sample sizes. So although the use of Cohen’s *d* might seem to be an arbitrary decision it does have some parallels to other areas of statistics. It is important to note that, although some justification exists for the use of Cohen’s *d*, not all researchers agree that statistical tests are unacceptable (e.g., Hansen 2008), however it would appear that a majority of researchers use Cohen’s *d*, and do so for the reasons stated here.

Inadequacy of current methods

Although there is a relatively clear method and background for assessing balance of means in propensity score analysis, we now come to one of the primary motivations for this paper: there are few methods for evaluating balance on higher order moments, despite methodological work and commentaries indicating that higher order moments can matter (e.g., Hill, 2008; Hill, Weiss, & Zhai, 2011; Sekhon 2007), and there is limited adoption of multivariate assessments of balance, despite the multivariate nature of the data. Sekhon (2007) gives a brief comparison of a few methods for comparing higher order moments, eQQ and Kolmogorov-Smirnov tests specifically, however these methods are not multivariate. Sekhon argues that multivariate methods would require extremely large samples. However, a failure to consider multivariate aspects, such as covariances, could still result in imbalanced samples, even in cases where univariate balance is perfect. It is easy to generate data for two groups with identical marginal distributions across an arbitrary number of covariates, and simultaneously giving the two groups completely different covariance matrices. For this kind of imbalance to have an effect all that needs to occur is for the treatment effect to be confounded with interactions of covariates. Assessing balance using only measures of the mean (the first moment) is not enough to state that the samples are adequately matched. Additionally, and relatedly, relying solely on univariate measures of balance is also inadequate.

The issues regarding the lack of multivariate testing deserves further elaboration. To account for the multivariate nature of data, and particularly interaction effects, some researchers have resorted to using Cohen's *d* to assess balance on interaction terms. However, this method raises another potential issue, because as the number of variables used in the propensity score increases the number of comparisons for the means and variances increases linearly whereas the number of comparisons for covariances increases quadratically (it follows the pattern of triangle numbers, specifically). This rapid growth of covariances greatly increases the burden on the researcher and increases the potential for false positives, especially if researchers need to check the balance of every interaction term. Of course, means, variances, and covariances are not the only moments that exist in data. Some existing methods, such as the cross-match statistic described below, attempt to solve both the problem of higher order moments and the multivariate nature of the data simultaneously.

Some attempts exist to assess more advanced methods in the context of propensity score analysis, but such attempts have generally been limited in their scope. Chen and Small (2016) compared the power of a few methods, some univariate and some multivariate (three of which are described in further detail below), to detect balance differences; however, their study was quite small with only 100 Monte Carlo replications in each of 8 conditions, all with multivariate normal data. Chen and Small also did not consider comparing multiple models fit to the same data, as might occur when a researcher is attempting to choose between multiple propensity score models. Their study is one of the most elaborate and comprehensive in this area, though quite limited. In another paper, Lee (2008), compared five different balance metrics. This paper is perhaps slightly more comprehensive than Chen and Small, however because of the age of the paper it has not compared newer metrics. Furthermore, Lee (2008) only used a single simulation condition based on an empirical study, which somewhat limits the generalizability of the findings. The current paper aims to improve on previous studies by using a greater number of metrics and a more diverse array of simulation conditions.

There are some possible reasons for the dearth of papers evaluating balancing metrics. The most obvious reason is that evaluating balancing methods is not entirely straightforward. Unlike evaluating modeling methods where there are clear benchmarks for performance (e.g., which model recovers the population parameters the best?) there are not necessarily directly equivalent benchmarks for balance metrics. Ideally one would use the metric that provides the most accurate assessment regarding which PS model produces the most balanced data; however, without knowing which balancing metric is the best it is difficult to determine which PS model produces the best balance, a catch-22. In the current paper I evaluate the balance metric on the ultimate performance of the PS model "chosen" by a given balance metric. There are multiple PS models compared, some will provide a less biased estimate of the treatment effect than others. The best balance metric ought to result in decisions that result in the least bias. While this is a practical method of evaluating the balance metrics, it could be criticized. Sekhon (2007) suggested that no balance metric is a monotonic function of bias, and so this evaluation technique may be criticized on those grounds. Sekhon is somewhat unclear, and so it is not obvious if he means that within a given sample the most balanced sample may not provide the most accurate result, or if he means that in the population the most balanced data may not have the lowest bias. The former result is an obvious conclusion that can be reached when considering the effect of random noise; a balanced dataset may not have the lowest error simply due to random chance. In the population it is difficult to see how this result would occur. Any systematic imbalance ought to introduce systematic bias into the results so in the limit we would expect that only the true propensity score model (or a model that provided equivalent balance) would have no bias.

Chapter 2

Methods

To begin the methods section I am outlining each of the three main “moving parts” in this simulation study: the conditions tested (and justifications for each), the models fit, and the methods assessed. The conditions are what determine the details of the data being analyzed. The models are, in this context, the propensity score models fit as well as any other means used to make the data comparable (e.g., randomization would be a “model” in this context). Finally, metrics are the means used to assess balance after matching. To summarize: a model is used to balance data and a metric is used to assess the degree of balance.

I will briefly summarize each of these three aspects here and describe them in greater detail below. Table 1 shows the considerations for each condition. I will use a factor structure to determine the correlations between variables. The covariance matrices will have as their baseline a given correlation matrix, but depending on condition may be altered by increasing the standard deviation. It is possible that the efficacy of some measures may be influenced by the number of variables observed or the sample size and so both of those are varied as well. Finally, the ability for a given method to detect differences in the shape of the distribution is important and two different distributions are used for the treated group, the control group data will always be multivariate normal.

Table 2 outlines the models compared. The models are how data are balanced. As part of the evaluation of the methods it is important to know how they would inform a researcher

Table 1: Conditions

Within Factor Correlations	0, .1, .3
Between Factor Correlations	0, .1, .3
Within Factor Correlation Differences: Treated vs. Control	-.3, -.1, 0, .1, .3
Standard Deviation Ratio: Treated/Control	2/1, 1/1, ½, 1/10
Variance in the Control Group (Predictors of Treatment)	1, 2
Number of Variables per Factor	1, 2, 3
Sample Size: Treated (Control)	50 (500), 250 (2500)
Treatment Effect Size	1
Treated Population Distribution	Multivariate Normal, Multivariate Log-Normal

attempting to choose between multiple models, some of which are correct in the population and some that are not. Propensity score models will use logistic regression and the MatchIt package in R to create matched samples of treated and control individuals.

Last are the metrics being compared. These are the means of assessing balance after the data are collected and matched. These metrics are listed in Table 3 along with a list of group differences that the metric could detect (feature). Note that just because a metric has a specific feature does not mean that the metric is particularly powerful in detecting that difference. Of primary interest in this study is the comparison between metrics that compare higher order moments and those that do not and between metrics that are multivariate and those that are not.

Table 2: Models

Original Trial	No balancing or attempts to equate groups. A completely naïve approach.
Bad Propensity Score Model	Includes only main effects and only for irrelevant variables.
Incorrect Propensity Score Model	Includes only main effects but for all relevant variables.
Correct Propensity Score Model	Includes relevant main effects, bivariate interaction terms, and quadratic terms.
Over Specified Propensity Score Model	Includes all main effects, bivariate interaction terms and quadratic terms for all variables related to either treatment or outcome.
Randomized Trial	Data that simulate the effect of random assignment

To summarize, the conditions determine the data that will be produced and analyzed. The models are used to create balance among covariates as would be expected in typical research settings. Multiple models are used for balance to simulate the condition where the researcher

Table 3: Balance metrics

Method	Means	Variance	Covariance	Distribution	Multivariate?	Does Sample Size factor

						into calculation?
Cohen's *d*	Yes	Not Directly	Not Directly	Not Directly	No	No
RMSEA	No	Yes	Yes	No	Yes	Yes (factored out)
Jennrich χ^2	No	Yes	Yes	No	Yes	No
Cross-match	Yes	Yes	Yes	Yes	Yes	Yes
Nearest Neighbor Cross-match	Yes	Yes	Yes	Yes	Yes	Yes
Minimum Spanning Tree Cross-match	Yes	Yes	Yes	Yes	Yes	Yes
RMR	No	Yes	Yes	No	Yes	No
Kolmogorov-Smirnov	Yes	Yes	No	Yes	No	No
Hotelling's T^2	Yes	No	No	No	No	Yes
Group Mean Mahalanobis Distance	Yes	Not Directly	Not Directly	No	Yes	No
d^2	Yes	No	No	No	No	Yes

does not know, a priori, what the correct propensity score model is. Finally, the metrics are what is of interest to this paper and are being compared to each other. Each metric will be applied to each model in each condition. Metrics will be evaluated on their performance in choosing the best models in each condition. Further description of these facets of the study follows.

Conditions

The basic scheme for data generation derives from similar schemes used in other simulation studies in the propensity score literature (e.g., Austin, Grootendorst & Anderson 2007). There will be four types of variables as shown in Table 4. True confounds are the variables that are most concerning from an analysis perspective. These are variables that influence both treatment assignment and the outcome variable. For example, a person's baseline depression level may affect their propensity to seek treatment as well as their post-treatment depression score. Stochastic confounds are variables that are associated with the outcome variable but not the propensity to receive treatment. If this variable is not balanced between the treatment and control groups it will affect the mean difference between the two groups, however it ought to vary between the two groups at only chance levels and the average difference between groups (in the population) is 0. Because the differences between the two groups due to this variable should be due only to random differences between the two groups we will call this kind of confound a stochastic confound. Continuing the depression example, perhaps the cause of depression (e.g., job loss vs. loss of a loved one) predicts the level of depression a person experiences but does not (in the population) predict whether they will seek treatment for their depression. Next in the table are false confounds. These are variables that do affect treatment assignment but are themselves entirely unrelated to the outcome variable. For depression, perhaps psychology students are more likely to seek treatment for depression but are no more likely than engineering students to experience depression. Lastly we have red herring variables. These are variables that predict neither treatment nor outcome. These variables ought to have no bearing on the actual outcome of the analysis (as the name implies). In the depression example this could occur if a researcher believed that smoking was related to depression but it was neither related to the propensity to obtain treatment nor was it related to depression itself. Naturally, all of these described associations are at the population level.

Table 4: Types of confounding variables

		Predicts Outcome?	
		Yes	No
Predicts Treatment?	Yes	True Confound	False Confound
	No	Stochastic	Red Herring

In many Monte Carlo simulations examining propensity scores, the propensity score is calculated and then used (typically in conjunction with a random number generator with a Bernoulli distribution) to assign treatment. The resulting treatment assignments define the treated and control subjects for a given simulation, a propensity score analysis is conducted, and the method of interest is examined. However, for the present study I propose a different approach. The primary concern that propensity score analysis is meant to rectify is that the underlying distributions of covariates for the treated and control subjects are different. A propensity score is meant to adjust for differences between the treated and control groups, but was not initially described in the literature as the data generating mechanism behind treatment and control group assignment. Instead of using the propensity score to generate differences in treated and control groups I will directly create the differences between the two groups. This approach has multiple benefits. First, it makes it analytically simpler to control differences between the groups. Rather than referring to values in the propensity score model and attempting to carefully assign those values to create distinct differences, I can simply create the differences. Second, I can control the sample sizes directly. In the commonly used method of simulation there are differences in the size of the treated and control groups in each simulation because treatment assignment relies on random observations of a Bernoulli variable. Using my method I can create exactly the number of treated and control subjects required in each condition.

For simplicity I propose using factor structures to determine the correlation matrices. These factors do not represent anything substantive, but are a tool to create correlation matrices that are (1) not sparse, (2) can be algorithmically generated, and (3) have multiple unique correlation values in each matrix (e.g., not all the correlations are .3). The factors will be based on the taxonomy of confounding variables presented above. For example, true confounds will form a single factor, stochastic confounds will form a separate factor, etc. Within a factor variables will be either not correlated (0), minimally correlated (.1), or moderately correlated (.3) with each other (three levels). The factors also will be allowed to have no, minimal, or moderate correlations with each other. These correlations have been chosen because they represent what a researcher may reasonably expect to see in psychological research. In order to allow for differences between groups that still result in plausible correlations for psychological research the baseline correlations were kept on the lower end of what might be expected in research settings.

To test the sensitivity of the method to differences between correlation/covariance matrices, and tangentially the ability of PS matching to correct for these differences, discrepancies between the correlation/covariance matrix for the pool of treated and the covariance matrix for the pool of control subjects will be introduced. These discrepancies will be limited to the true and false confounds, i.e., these discrepancies will occur only in the variables that predict treatment status. Discrepancies will be introduced in the loadings by adjusting the correlation matrix for the treated subjects. The adjustment will be either small ± .1 or medium ± .3, (or the baseline condition of 0, no difference). For example, if in the control group the correlation between two variables in the same factor is .3 and the adjustment is .3, then in the treated group the correlation between those variables will be .6. These differences were chosen for their symmetry with the baseline correlations and because the maximum possible correlation, .6, was thought to be within what might be routinely observed in research settings, but at the higher end of what a researcher might expect to observe.

So that the results are generalizable beyond differences in correlations, the standard deviation of the true confounds will be allowed to vary in the treated group. The standard deviation of a truncated normal distribution can be calculated using formulas found in Barr and Sherrill (1999). If one creates a cut score such that only the top 1% of the population is beyond the cut point, that truncated distribution has a variance approximately 1/10th that of the parent distribution. Using a standard deviation ratio, instead of a variance ratio, of 1/10th creates an extreme condition that ought to be more stringent than even the most unusual of real world cases. I will allow the variances of the treated group to be either twice, equal, ½, or 1/10th the standard deviation in the control group, these differences in standard deviations will be introduced only in the variables that are different between groups (i.e., the true and false confounds). The control group variances will be either 1 or 2 (standard deviations of 1 and $\sqrt{2}$). So, if the standard deviation of a true confound in the control group is 1, and the ratio between the treated and control groups is 1/2, the standard deviation on that same variable in the treated group will be set to 1/2. The distribution of differences is asymmetric around 1, the primary motivation for this is that it seems more plausible for a highly select sample to have restriction of range (and therefore standard deviation) relative to the general population, so in general it seems more likely that the group receiving treatment would be the group with generally lower variability. There is an additional, practical, reason to not have perfect symmetry in this case. Because the treated are being matched to the controls, and because there are many more controls than treated, if the treated have a smaller standard deviation it is generally still possible to find adequate matches. If the treated have a substantially higher standard deviation than the control subjects then many of the treated subjects will go unmatched since they will be outside the range of data for the control subjects.

The four-category taxonomy for confounds makes it convenient to increase the number of observed variables in increments of four. For the present analysis we will use 4, 8 and 12 predictor variables. The use of multiple numbers of variables should help reveal if the metrics for comparing covariance matrices are affected by the number of variables. As an a priori consideration, it seems reasonable that with larger covariance matrices any metric would be more unstable as the number of observations is fixed. Although any one covariance or variance in a given matrix will be estimated at the same precision as any other measured using the same sample size (and same measurement reliability, etc.) two 12 X 12 covariance matrices simply have more opportunities for a large difference between the same position cells than two 4 X 4 matrices. There is a practical reason for limiting the number of variables to a maximum of 12, in the subsequent section on models 12 variables results in models that are at the limits of estimability. Increasing the number of variables would necessitate increasing the lowest sample size somewhat dramatically.

Two sample sizes will be used. I propose using a sample of 250 treated and 2,500 controls for a “large” condition representing a relatively large study, and 50 treated with 500 controls for a moderate size study condition. The larger size is intended to replicate what might be expected from a well-funded research study, while the smaller size is what might be expected from a more convenience-based approach. Those sizes will result in matched sample sizes of 500 and 100. For some metrics, and for matching generally, the time increase involved in including larger samples is non-linear, for example, a single condition for the 50 and 500 sample takes approximately 10 seconds, while the same condition with the sample of 250 and 2500 takes approximately 10 minutes. The proposed sample sizes reflect what might be expected from a typical research study while still limiting the overall length of the simulation.

There will be two levels of the “distribution condition,” defining the distributions used to generate data for the treated groups -- a multivariate normal distribution and a multivariate log-normal distribution. For the control group only the multivariate normal distribution will be used. The normal distribution represents a best case scenario. If the method is unsuccessful in the case of a normal distribution then the method is unlikely to work for other distributions. A multivariate log normal distribution will be used as it has a substantial skew. In order to create this distribution uncorrelated normal deviates will be generated such that they have a population variance of 1 when exponentiated. These deviates will be grand mean centered on the population mean so that the population mean is 0. Then, using a diagonalization method from Kaiser and Dickman (1962) the desired covariance matrix is imposed on the generated data. The control group will always have a multivariate normal distribution. Although the normal and log-normal distributions hardly cover the full range of possible distributions they ought to offer insight into the viability of the methods used here under both optimal and non-optimal conditions.

In addition to confounds, the outcome variable needs to be defined. The outcome will be the sum of the unit weighted true confounds and stochastic confounds, unit weighted quadratic terms of the true confounds, unit weighted bivariate interactions of true confounds with other true confounds (if appropriate) and the treatment effect of 1. The bivariate interactions and quadratic terms are omitted for the stochastic confounds because the primary concern is modeling the true confounds, and including these terms for the stochastic confounds would unnecessarily increase the complexity of the study. In sum (Table 1) there will be one effect size, two sample sizes, two population distributions, three different covariance matrix sizes, three different base loadings between variables in a factor (including 0), three different correlations between factors, two different variance levels, five differences between loadings (including 0), and four different variance ratios between treated and control. The mean values will also differ, by half a standard deviation, between treated and control subjects on variables predictive of treatment. As the focus of this study is not on balance of means the difference in means will be constant across conditions. This design results in 4,320 different conditions to test.

Models

Six “models” will be compared, four of which are actual propensity score models (Table 2). The first “model” is simply the original unmatched data. This will give a baseline for comparison. Next is the “bad” propensity score model, which will use the main effects of only the red-herring variables. This model is a worst case scenario. The next model is the incorrect propensity score model which will use only the main effects that differ between groups. This model is much better than the “bad” model; however, it omits key effects. The next model is the correct propensity score model and includes the main effects, interaction effects and quadratic effects that vary between the treated and control groups (i.e., for true and false confounds). This model includes all effects that vary between the treated and the control group in the population. It is the best model (in theory) that could be constructed without reference to what variables predict the outcome. The third model is an overspecified model, which will include the main effects, quadratic effects and bivariate interaction effects for all variables that have an effect on either treatment status or outcome. In order to keep the number of parameters in the overspecified model to a number that can be fit with the smallest sample size I will limit interaction effects to bivariate interactions between variables in a factor. In the largest model this will result in 36 parameters. Without this constraint there would be 63 slopes in the largest model (i.e., if all bivariate interactions were included). A final comparison “model”, a randomized trial, will be simulated by creating an additional control group with the same population means and covariances as the treated group. This final case serves as a “best case” scenario for the comparison of two distinct groups and would be at the upper end of performance expected from any of the propensity score models. I will use nearest neighbor matching (without a caliper) for simplicity, this will maintain sample sizes across all models within a condition.

Balance Metrics

There are several metrics available to assess balance on more than simply the univariate means. Table 3 lists all the metrics that will be evaluated in this paper. For covariances there are several metrics. Jennrich (1970) used matrix determinants and the average of two covariance matrices to calculate a χ^2 statistic. Using Jennrich's (1970) statistic it is possible to calculate an RMSEA metric (the sample sizes, degrees of freedom for the test and the χ^2 statistic are all available). Another approach is based on the Root Mean Square Residual (the RMR, the unstandardized version of the SRMR). In addition there are metrics that exist to compare the entire multivariate distribution. One such metric, developed by Rosenbaum (2005), is distribution free and exact, and has seen use in some propensity score settings (e.g., Heller, Rosenbaum, & Small, 2010). Chen and Small (2016) made some improvements to Rosenbaum's statistic. Chen and Small compared their metrics to Hotelling's T^2 and so I include Hotelling's T^2 as well. Diamond and Sekhon (2013) and Sekhon (2007) state that Kolmogorov-Smirnov tests form the basis of their "genetic matching" algorithm, an algorithm that creates a matching model on the basis of optimizing a balance metric. Since the K-S test is the default for some algorithms, and because it is a means of comparing higher order moments, I include it in my analyses. Two other metrics that I also will evaluate are somewhat more ad hoc. The first is to sum the absolute value of the Cohen's *d* statistic for all variables, their quadratic terms, and the interaction terms. This metric is conceptually similar to what researchers frequently do (e.g., Austin 2008). The primary difference between this metric and the common practice of comparing Cohen's *d*s is that I am summing all the individual statistics to give a single statistic. This is to allow for an automated decision to be made in each of the Monte Carlo trials. Another alternative metric is to calculate the Mahalanobis distance of the matched control means from the treated means. This metric may only assess balance on the means, but has the potential advantage of being an omnibus test. Lastly, I include d^2, an omnibus measure introduced by Hansen and Bowers (2008). They describe it as a "first cousin" of Hottelling's T^2.

Two potential tests were rejected on the basis of being either unfeasible in the present study, or mathematically similar to tests already included. The Least Squares Density Difference (LSDD) estimator is a promising means of evaluating the similarity of multivariate distributions (Sugiyama et al., 2013), however it requires a cross-validation procedure to pick necessary parameters. The iterative procedure used to calculated the LSDD would dramatically increase the simulation time (potentially adding months to the simulation). Another metric, the Multivariate Kolmogorov-Smirnov (e.g., Friedman & Rafsky, 1979) test is mathematically quite similar to the MST cross-match statistic presented by Chen and Small (2016), at least as implemented in available software (e.g., the GSAR package; Rahmatallah, Zybailov, Emmert-Streib & Glazko, 2017). The metrics now will be defined in greater detail.

The first metric is the covariance comparison approach by Jennrich (1970). Similar to deviance statistics in SEM, this metric compares two covariance matrices and produces a χ^2 test statistic which is asymptotically χ^2 distributed only if the underlying data are multivariate normal in each of the two groups). For statistical tests this could be problematic; however, for PS analysis, as long as matched samples with better balance have lower statistics, we can likely relax the assumption of multivariate normality. This approach only solves two parts of our problem. It provides a way to compare covariances, and it is an omnibus test so the issue of rapidly increasing numbers of tests is obviated. However, the use of χ^2 statistics causes certain issues which should be familiar to anyone acquainted with SEM. As sample size increases, the power of this test also increases, so even small differences (that a researcher may consider meaningless) can cause this test to be statistically significant. There are two potential solutions, and I test both here. The first is to simply calculate the raw test statistic and ignore the *p* value. The second potential solution is to use SEM fit indices (here I will use RMSEA; Steiger & Lind,

1980) to compare balance across-matched groups in propensity score analysis. RMSEA requires a deviance function with a χ^2 distribution, for which Jennrich's statistic may be suitable.

The exact formula for the deviance statistic illustrated in Jennrich (1970) follows. The χ^2 deviance statistic takes the form of $\frac{1}{2}tr(Z^2)$ where $Z = \sqrt{\frac{n_1-1}{n_2-1}} \times R^{-1}(S_1 - S_2)$, R is the average of the two covariance matrices S_1 and S_2 weighted by the sample size of each group minus 1. The degrees of freedom for this test are $\frac{p\times(p+1)}{2}$, the number of unique elements in the covariance matrix. The proof that this statistic is χ^2 distributed relies on the data having a multivariate normal distribution.

The second metric is based on the RMR, another statistic from the SEM literature. It is a conceptually simple statistic. To calculate the RMR the difference between each cell in the lower triangle (including the diagonal) of the covariance matrices from the treated and control group is squared and then the square root of the mean of those values gives the RMR. If the covariance matrices are equal between the treated and matched control groups then the statistic is 0, it increases as differences occur (and increase) between the two groups. Some research (Maydeau-Olivares & Shi, 2019) suggests that the standardized counterpart of RMR, the SRMR, may be better than other fit indices, such as RMSEA, in SEM settings. It is possible that it will have similar properties here.

The third metric, the cross-match statistic developed by Rosenbaum (2005), is more advanced than others, but still straightforward. Two samples of equal size are created, either via sampling or matching. The samples are combined and rematched using optimal non-bipartite matching, which matches subjects to their closest counterpart regardless of group membership in such a way as to minimize an overall distance statistic for the final matched sample. The researcher then calculates the number of outgroup matchings (i.e., cases where a treated is matched with a control). This number follows an exact distribution which allows for *p* values to be calculated. This metric makes no assumptions regarding the distribution of the underlying data and thus is very flexible. Rosenbaum's original work suggested that it was powerful for some differences and not for others. Power appeared to be low particularly when distributions were the same (e.g., both were normal) but had moderate differences in location or spread. This initial work was based on bivariate data. Chen and Small (2016) recently presented two similar statistics. The first is based on the minimum spanning tree of the treated and control units. The second metric is based on the nearest neighbor information of the treated and control groups. These statistics, like Rosenbaum's, evaluate how many within vs. between matches there are for the treated and control groups. Chen and Small also correct their *p* values to account for certain pathological cases. They concluded in their study that their metric has generally good power to detect imbalance, and that this power is better than that for Rosenbaum's cross-match statistic. As noted previously, they examined a relatively limited number of conditions. In the present study I compare Rosenbaum's cross-match statistic, without calculating a *p* value (i.e., using the raw statistic) as well as the *p* values from the two metrics proposed by Chen and Small (2016).

Hotelling's T^2 is a metric that compares multiple means in up to two samples. It is a generalization of Student's *t* statistic (Hotelling, 1931). The statistic is implemented in the R package "Hotelling" (Curran, 2018). The formula, assuming equal variances in the population, is $\frac{n_1 \times n_2}{n_1 + n_2} \times (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$ where S^{-1} is the pooled (averaged) covariance matrix for the two samples, $\bar{x}_1$ and $\bar{x}_2$ are the respective vectors of variable means for the two samples, and n_1 and n_2 are the respective sample sizes of the two groups.

The Kolmogorov-Smirnov (K-S) test is a common test for assessing distributional similarity and is implemented in the R package "stats" (R Core Team, 2019). The K-S test is also the default balance statistic optimized in the genetic matching algorithm provided by Diamond and Sekhon (2013). Given its status as a default in some software (e.g., Sekhon's "Matching" package in R; Sekhon, 2011), and that it is a popular method for comparing two univariate distributions it is included in the present analysis.

For the purposes of the current simulation I am including two additional statistics. The first calculates Cohen's *d* for all the variables, their quadratic terms, and their bivariate interaction terms. It then calculates the sum of the absolute value of all the Cohen's *d*s. This statistic is conceptually similar to the current most common practice. It differs in that it sums the statistics to create an omnibus statistic, and it includes terms regardless of their inclusion or exclusion in the PS model. The other statistic calculates the Mahalanobis distance of the variable means in the control group from the treated group. This statistic compares only group mean; however, it utilizes an existing distance metric, and that metric does utilize information about the covariance matrix in its calculation. The calculation for the Mahalanobis distance explicitly includes the covariance matrix of the group to which a comparison is being made. In the present analysis the statistic takes the form $\sqrt{(\bar{x}_{control} - \bar{x}_{treated})^T S_{treated}^{-1} (\bar{x}_{control} - \bar{x}_{treated})}$. The covariance of the treated group is present in the inverse form $S_{treated}^{-1}$, but the metric itself computes the multivariate distance between the means of the treated and control groups. This statistic is almost a scaled version of Hotelling's T^2 except that it does not use the pooled covariance matrices.

The metric d^2 is another omnibus measure. It is described as being a first cousin of Hottelling's T^2 by its creators (Hansen & Bowers, 2008). It was originally designed with cluster randomized trials in mind, but it is easily adapted to propensity score analysis, and in fact it is a suggested application by Hansen and Bowers. It is implemented in *R* in the package RItools using the function xBalance(). Mathematically it appears to create cluster weighted group mean differences and then combine them into a single omnibus test in a way similar to Hottelling's T^2. In the present application it seems somewhat unlikely that it will have much advantage over T^2.

Chapter 3

Results

Within each iteration of all 1,000 replications in all conditions all balance metrics were calculated for each of the models, namely the bad PS, the incorrect PS, the correct PS, the overspecified PS, and the randomized trial. For the unmatched original data, a random subset of the control subjects, equal in size to the treated sample, was taken. Balance metrics for the unmatched data were calculated on this reduced sample. It was necessary to use a reduced sample because some of the balance metrics become exponentially slower with increasing sample sizes. For a given simulation cycle in a given condition the most important data were: the values of the balance statistics for each model and the estimated treatment effect of each model. From this I was able to "select" a model in each cycle using each metric, and then calculate what the overall error would be for a given metric in a given condition. A model was selected using a balance metric if it had the best value on the given balance statistic. Error was defined as the square of the difference between the observed treatment score and the population treatment score, here abbreviated SQE (SQuared Error) and assigned that SQE to that metric for that condition. I also calculated the SQE for each propensity score model, the randomized trial and the unmatched data.

Table 5: Mean squared error of each model over all conditions

Bad PS	Incorrect PS	Correct PS	Overspecified PS	Randomized Trial	Unmatched Data
216.42	174.43	167.14	171.96	146.44	216.38

As a first check it would be useful to know if the “manipulation” worked. Here the manipulation is the different propensity score models. Ideally there would be a gradation of performance (as measured by SQE) and in a somewhat logical pattern. Ideally the pattern would be (from worst to best): Unmatched data, the bad PS model, the incorrect PS model, the correct PS model, the overspecified PS model, and the randomized trial. The position of the overspecified PS model might not be so fixed since it is technically an incorrect specification. All that said, so long as there is some variability in PS models the manipulation ought to be adequate. Table 5 shows the mean SQE across all cycles in all conditions of the analysis. In general the results follow the hoped-for pattern. The only reversal is slight, with the overspecified PS model performing slightly worse overall than the correct PS model. This result suggests that the manipulation had the intended effect and that there is variability in the effectiveness of the propensity score models that, hopefully, some of the metrics will be able to pick up on.

Aggregate results for the balance statistics are presented in Table 6. The metrics are evaluated by the mean SQE that would result if a researcher used that metric to choose which propensity score model to use. This table is limited to the results of propensity score models. Including the randomized trial and the unmatched data in this table would, under some conditions, produce extreme values which would likely make it easier for metrics to “pick” the correct model. Including the randomized trial and the unmatched data would inflate the performance of some metrics. The table shows the percentage of times, across all conditions, that a given metric had the lowest mean SQE out of all metrics being compared. In some circumstances some metrics consistently agreed on which model to pick, resulting in overlapping

Table 6: Percent of time metric had the best performance

	Jennrich χ^2	RMSEA	RMR	Mahalanobis	Aggregate Cohen's *d*	K-S Mean
Raw	7.87%	7.59%	41.94%	2.52%	6.67%	2.18%
Within 10%	53.15%	52.66%	56.20%	50.30%	50.88%	43.15%
	K-S median	T^2	d^2 χ^2 statistic	Cross-Match	Nearest Neighbor CM	*p*- value NNCM
Raw	0.90%	10.35%	10.35%	0.00%	6.37%	0.00%
Within 10%	38.31%	47.41%	47.41%	22.64%	29.91%	20.93%
	MST CM	*p*-value MSTCM	*p*-value Jennrich	Mean KS *p*-value	Median KS *p*-value	
Raw	6.97%	0.00%	3.73%	10.09%	0.21%	
Within 10%	29.61%	25.76%	22.18%	31.74%	23.77%	

Note: Raw gives the percentage of times that a given metric gave the best performance as measured by mean squared error, over all conditions. The "Within 10%" rows give the percentage of times that a given metric had a mean squared error within 10% of the best performing metric across all conditions. Performance was determined by the squared error of the models picked by a given metric. In the event that a metric would have picked multiple models their squared error was averaged.

results, therefore the numbers in Table 6 do not sum to 100% (as might be expected). I have also included in Table 6 an equivalent analysis looking at instances where a given metric's mean SQE was within 10% of the best metric's mean SQE. The headline result from this table is that the

RMR massively outperformed all other metrics when looking at absolute best performance. In nearly 42% of cases the RMR had the best performance, the next best performing metrics were T^2 and d^2. As a brief aside, further investigation revealed T^2 and d^2 to be behaving identically. In this simulation there was no case where I found any difference between the two measures. They appear to be mathematically equivalent in the present design. The reason for this equivalence is that the d^2 measure uses a weighting scheme, however in the present study with 1:1 matching it appears that the weighting results in a simple transformation of T^2. The more advanced cross-match statistics behaved surprisingly poorly in the present study. This result may be due to a

Table 7: Average percent of times, across conditions, that the metric picked the model with the least error in a given simulation.

	Jennrich χ^2	RMSEA	RMR	Mahalanobis	Aggregate Cohen's *d*	K-S Mean
Raw	32.39%	35.88%	35.13%	27.75%	27.94%	26.41%
	K-S median	T^2	d^2 χ^2 statistic	Cross-Match	Nearest Neighbor CM	*p*- value NNCM
Raw	29.92%	27.20%	27.20%	31.60%	31.92%	36.87%
	MST CM	*p*-value MSTCM	*p*-value Jennrich	Mean KS *p*-value	Median KS *p*-value	
Raw	29.33%	52.50%	59.19%	30.31%	31.59%	

Note: The numbers in this table are rather generous due to how ties are handled. If a metric in a given cycle of a given condition gave three identical responses for three different model (one of which was the correct model) and these responses were all the best value given by the metric in that cycle it was counted as "picking" the best metric despite the ambiguity.

relatively high number of ties for the cross-match statistics. When a metric would choose multiple “best” models it was necessary to account for this. In order to do so I averaged the SQE of the chosen models. Taking the average was a means of penalizing metrics that would give ambiguous results. This can be seen in table 7 as well. Table 7 is similar to Table 6 in that it is still evaluating how well metrics did at picking the best model. The difference is that this table shows the percentage of the time that a metric would choose the best performing model. Because of how ties are handled, if the metric gave a tied result (and the best result) to both the best performing model and some other model(s) the metric was counted as having picked the correct model. This is a generous evaluation of metric performance, but in conjunction with table 6 it reveals that the cause of the poor performance of the cross-match type statistics is not due to an inability to give a low score to the best propensity score model, but due to an inability to distinguish between the other propensity score models. It should also be noted that *p*-value statistics were especially prone to ties given that the software generally reported *p*-values to limited (e.g., 6) decimal places. In cases where the *p*-values were very large or very small this led to apparent ties. Fortunately, for every *p*-value statistic there is an equivalent raw statistic, and the raw statistics (e.g., the Jennrich χ^2 and the cross-match statistic) were not limited by the printing of decimal places, which allows a more direct evaluation of the performance of those metrics.

Moving on from aggregate statistics I thought it would be illuminating to see if it was possible to predict how well a metric performed. For this evaluation I will focus on the RMR as it was the best performing metric. The following analysis will be an exploratory data analysis. Fortunately there are pilot data available for a few thousand conditions, which will make for a good cross-validation sample. I will also define performance somewhat differently than in the previous analyses. Performance will be the difference between the RMR's SQE in a given condition, and the best possible SQE. The best possible SQE is the SQE that would result if a researcher knew exactly which model was the best choice in a given cycle. The closer the RMR SQE is to this best possible SQE, the better the performance. The RMR SQE minus the best possible SQE can never be negative. It might seem odd to use this metric instead of the ratio of the two. However, the ratio had the unfortunate property that it would become very large when differences were small. This approach creates a metric in which small deviations from the truth are penalized more heavily than steep departures. Because the same effect size was used in all conditions, a given difference in SQE retains the same meaning across all conditions (i.e., a single point increase is indicative of an average difference equal to the original effect size of one). In the following regression analyses a greater predicted value is a prediction of worse performance. The mean SQE difference aggregating across all conditions was 8.47. This is slightly more than the difference between the mean SQE's for the correct PS model and the incorrect PS model, but substantially less than the difference between the SQE's for the correct PS model and the bad PS model, or even the difference between the SQE's for the correct PS model and the randomized trial. First, I plotted the difference in SQE by condition; see Figure 1. The conditions were created sequentially and systematically, so a non-random pattern in the figure indicates a relationship between the SQE difference and the simulation conditions. It should be immediately apparent that there is structure to this data, which also indicates that it ought to be possible to predict performance.

The first step was to fit a regression model with main effects for the condition variables (e.g., the ratio of control to treated standard deviation). All condition variables were entered in their raw form except for the ratio of control to treated standard deviation. For this variable, the absolute value of the log of the variable was taken. This transformation gives the same value when the ratio between the treated and control standard deviations is 2 as when that same ratio is ½ (i.e., if one group has a standard deviation double that of the other, then the value is the same regardless of which group it is). This variable will be referred to as the ALRSD (absolute log ratio of the standard deviations) going forward. The model R^2 was .22 and was statistically significant ($F_{8,4311} = 153.7$, $p < .001$). All main effects were statistically significant except for the baseline correlations and the between factor correlations. Table 8 presents the results of this first model. Increasing the number of variables reduced performance of the RMR, as did increasing the baseline standard deviation and moving from normally distributed data to log-normally distributed data. A higher ALRSD increased performance, as did an increase in the value of the discrepancy of the treatment group within factor correlations relative to the control group within factor correlations. This last find is a bit surprising. If this value were -.3, for example, it would

Figure 1: Plot of RMR SQE minus the best achievable SQE by condition

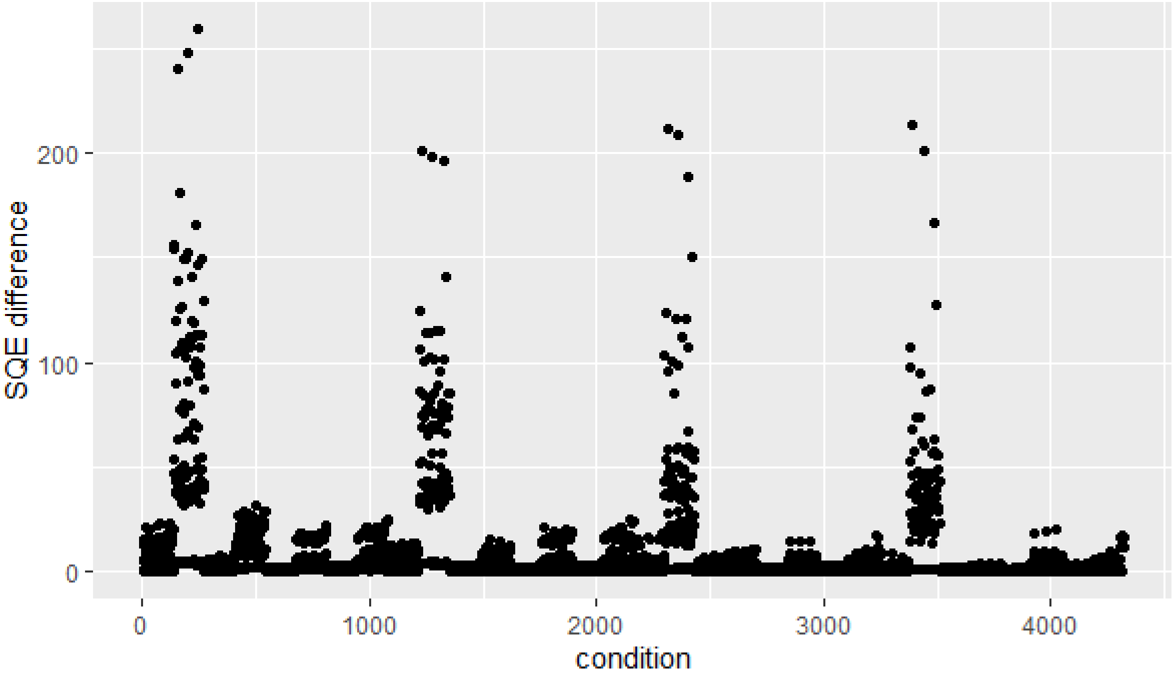
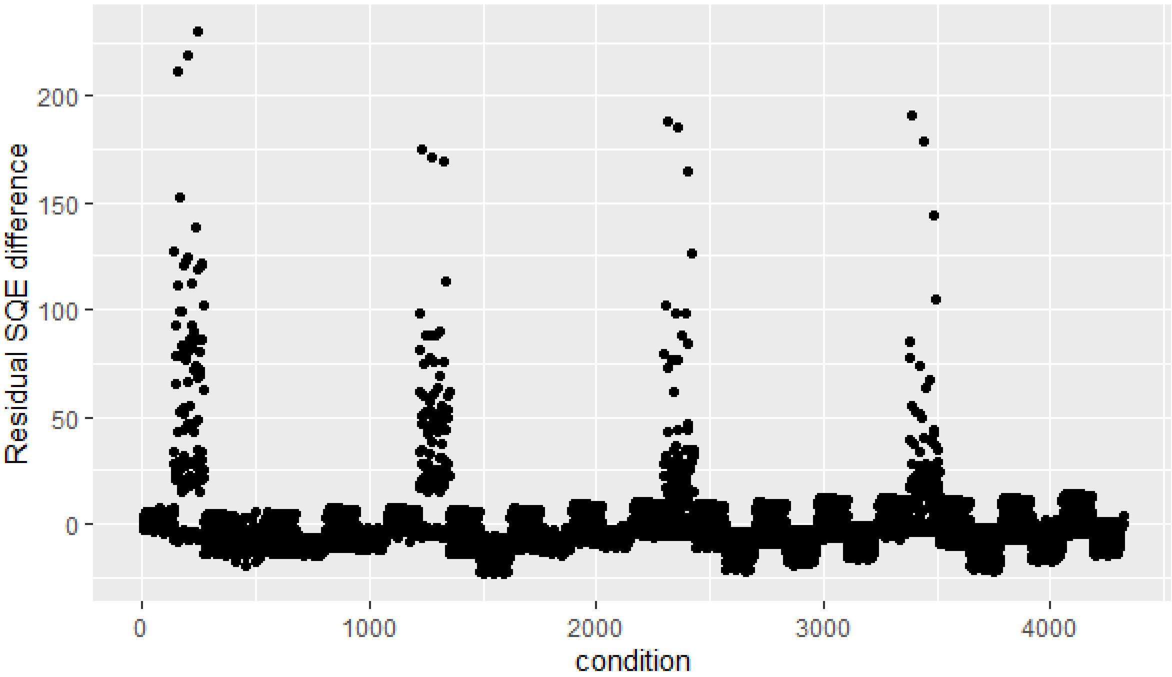


Table 8: Results of first regression model

	Estimate	Standard Error	t value	*p*-value	
Intercept	-24.09	1.35	-17.91	0.00	***
Number of variables per factor	8.38	0.35	24.24	0.00	***
Within factor baseline correlations	0.88	2.26	0.39	0.70	
Between factor correlations	1.60	2.26	0.71	0.48	
Baseline Standard deviation	12.86	0.56	22.78	0.00	***
ALRSD	-1.26	0.33	-3.79	0.00	***
Treatment group deviation from baseline within factor correlation	-6.61	1.41	-4.68	0.00	***
Distribution (1 is log-normal, 0 is normal)	1.99	0.56	3.52	0.00	***
Sample Size	-0.02	0.00	-8.59	0.00	***

suggest a better performance of the RMR. However, if the baseline correlation were .3, then this is a lower correlation in both value and magnitude for the treated group compared to the control group, but if the baseline correlation were 0, then this means the treated group has a lower correlation in value, but not magnitude, relative to the control group. This result implies

Figure 2: Plot of the residual RMR SQE minus the best achievable SQE by condition for model 1



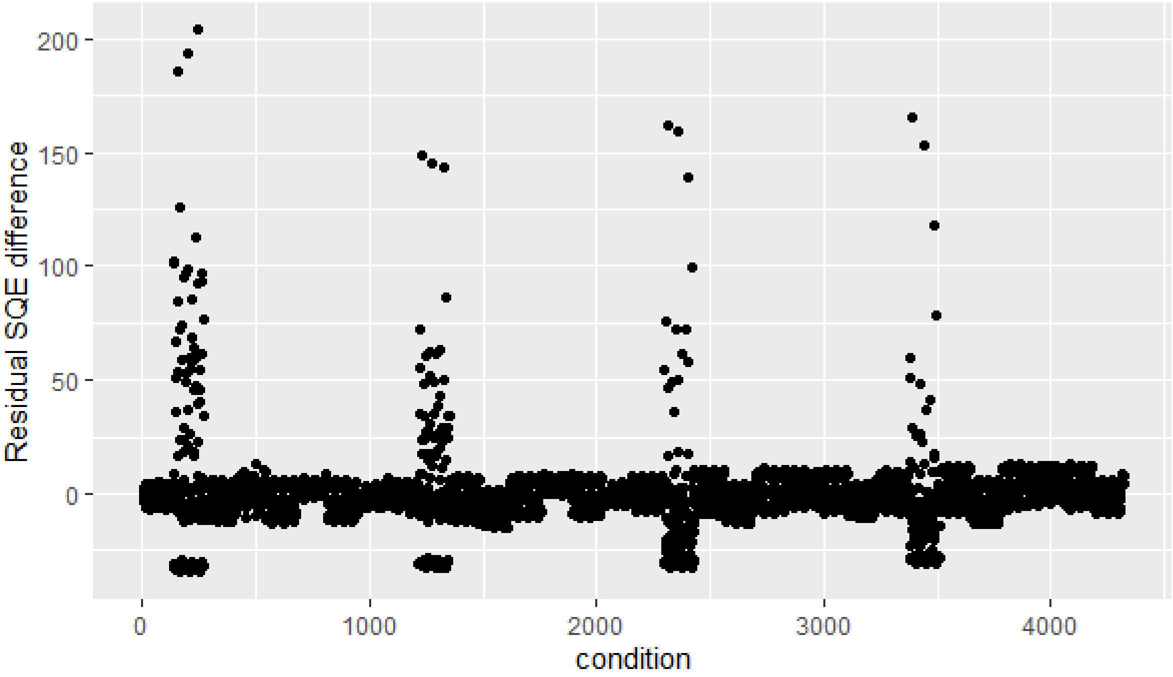
that the sign of the correlation matters, but there is not obvious logical reason for this paradoxical finding, which will be investigated in a subsequent analysis. Looking at the residuals we obtained in Figure 2, there are two clear conclusions that can be drawn from this plot. First, there are four clusters that are both heteroscedastic and are likely contributing to much of the residual variance. Second, there is still systematic variability, made apparent by the “blocking” of the residuals.

The second model fit attempts to solve the mystery of why the sign of the correlation difference seems to matter. The first model in this attempt removed the within factor baseline correlation. This removal had no effect on model fit. Next, I calculated what the within factor correlation would be for the treated group in each condition and substituted that for the currently included difference between treated and control correlations. This adjustment resulted in a statistically significant reduction in fit, which arguably should have had no effect since it is a linear combination of two variables already in the model, one of which had no effect on the model. Next I took the absolute value of the difference. This gave a better R^2 than the previous model, but not as good as the original fit. Another model included the absolute value of both the difference between the treated and control correlations and the absolute value of the treated correlation. This model had slightly worse performance than the first model (and R^2 =0.2207 vs. an R^2 = 0.2219), however the parameters are, in my opinion, more interpretable. In this model an increase in the magnitude of the difference between treated and control group correlations improved the performance of the RMR, but a simple increase in the magnitude of the treated group correlations decreased performance.

Moving on from the effect of correlation differences I next examined the impact of when the covariance matrix of the treated group, control group, or both was identity. The main effects had no impact. The interaction effect was marginally statistically significant (p = .04), but that hardly seems meaningful in this context. A plot of the residuals revealed no substantial changes. These terms were dropped from subsequent models.

Next I added an interaction between the baseline standard deviations and the ALRSD, this caused a modest increase in R^2. However, a typo in the first run of the model used the raw ratio of the standard deviations. This resulted in including both the main effect and the interaction with the baseline standard deviation in a model. That model more than doubled the R^2 to 0.4704. That is too big of an improvement to ignore, despite the accidental nature of the discovery. Including both the ALRSD and the raw variable and the interaction effects gave an R^2 of .4955. Reflecting on this finding, it may reflect a specific aspect of how matching works. Matches are from the treated to the controls. When the treated have a standard deviation greater than the controls it can result in there being few good matches for the treated, however when the controls have a greater variability than the treated, there are so many controls that it is still possible to find good matches for the treated. This problem is significant enough that a planned simulation condition where the treated would have had a standard deviation 10 times that of the controls had to be dropped due to persistent model failures. In the event that good matches are hard to come by it may be that small differences in balance cause large differences in the SQE. This pattern could result in the RMR choosing incorrectly between two PS models with similar balance but a large difference in outcome, inflating the RMR SQE relative to the best possible SQE. The residuals for the full model are shown in Figure 3.

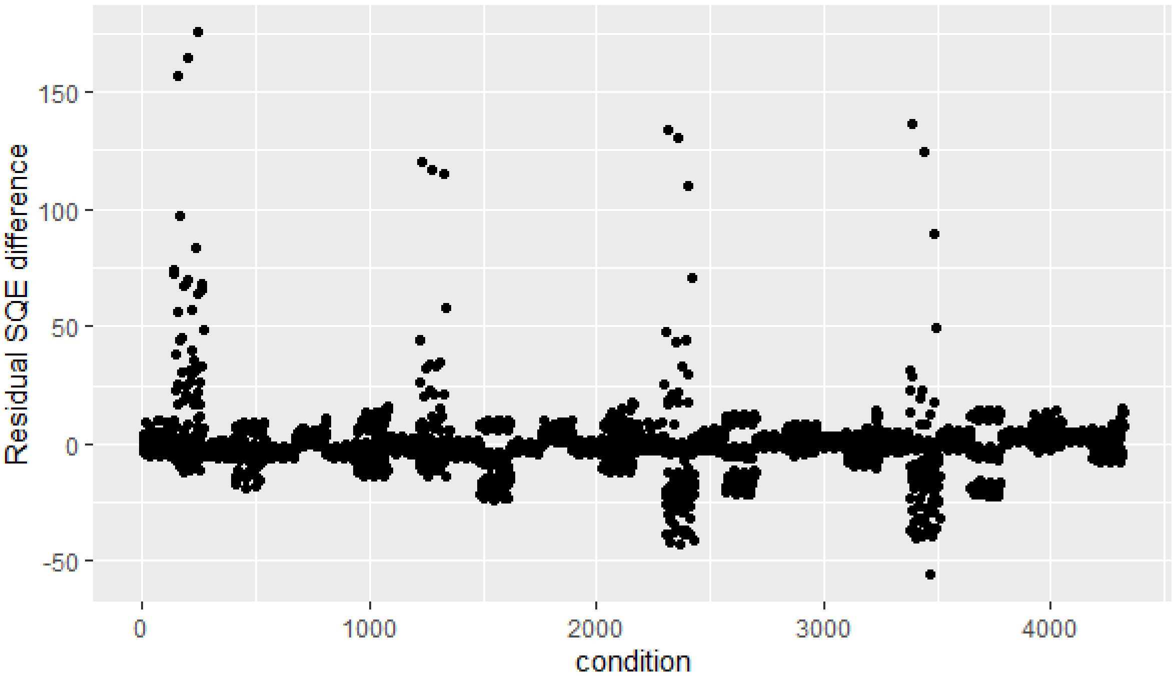
Figure 3: Plot of residual RMR SQE minus the best achievable SQE by condition for model 9



The next stage in the modeling was to find what was causing the four clusters of clearly heteroscedastic residuals. As a simple method I looked at the conditions that were applied in the 200th, 1,300th, 2,400th and 3,450th condition, conditions which lie in the middle of these clusters.

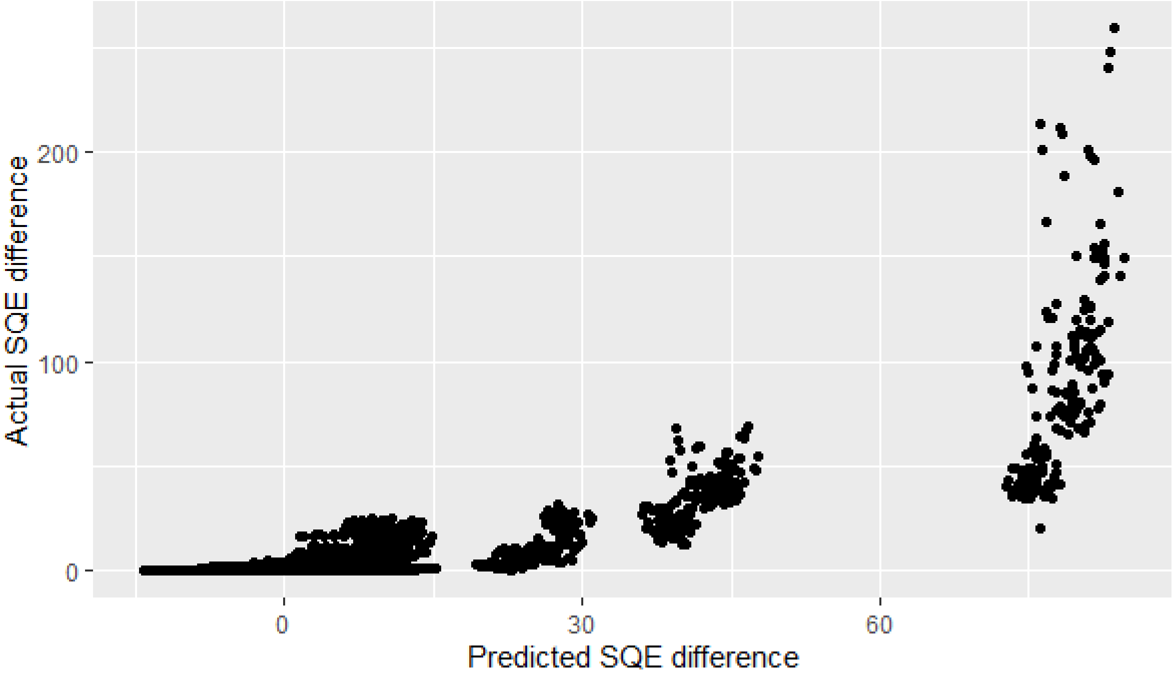
There were two common features of these four conditions. The correlation between factors was always .3, and the baseline standard deviation was always 1. Expanding the view to include a few observations before and after these selected observations showed that the baseline standard deviation remained the same, however the between factor correlation was not constant. Plotting

Figure 4: Plot of residual RMR SQE minus the best achievable SQE for model 10



the residuals excluding observations where the baseline standard deviation is 1 however showed that this was not the cause. This result however suggested an alternative method: plotting residuals against each condition variable. The variance of the residuals was much higher in conditions with 3 measured variables per factor, it appeared to decrease as correlation differences increased in value (but not magnitude), was substantially higher when the baseline standard deviation was two, and was also much higher when the ratio of the standard deviations was 2. I added an interaction between the number of variables in a factor, the standard deviation and the ratio of the standard deviations to the model. This adjustment increased the R^2 to 0.7109 from the previous 0.4955. As these models were nested it was possible to do a model comparison. The full model fit better *F* = 1,070.9, *p* < .001. The plot of the residuals is Figure 4. There still exist the four distinct clusters of heteroscedasticity, however the range of these clusters has shrunk. This figure also makes it possible to see that there is still some structure to the residuals. Looking again at plots of the residuals against condition variables shows that much of the previously noted relationships between heteroscedasticity and the variables remains.

Figure 5: Plot of RMR SQE minus the best achievable SQE by the predicted SQE difference for model 10



Trying a different tactic I plotted the predicted values from the model against the observed values, which produced figure 5. Figure 5 is helpful in that it suggests looking at the conditions where the predicted difference in SQEs is greater than 60. In all those conditions there were three variables per factor, the baseline standard deviation was 2 and the ratio of the standard deviations was 2 (with the treated group have the greater standard deviation). These plots suggest a data transformation. In retrospect this perhaps should have been obvious given

Figure 6: Histogram of raw SQE differences

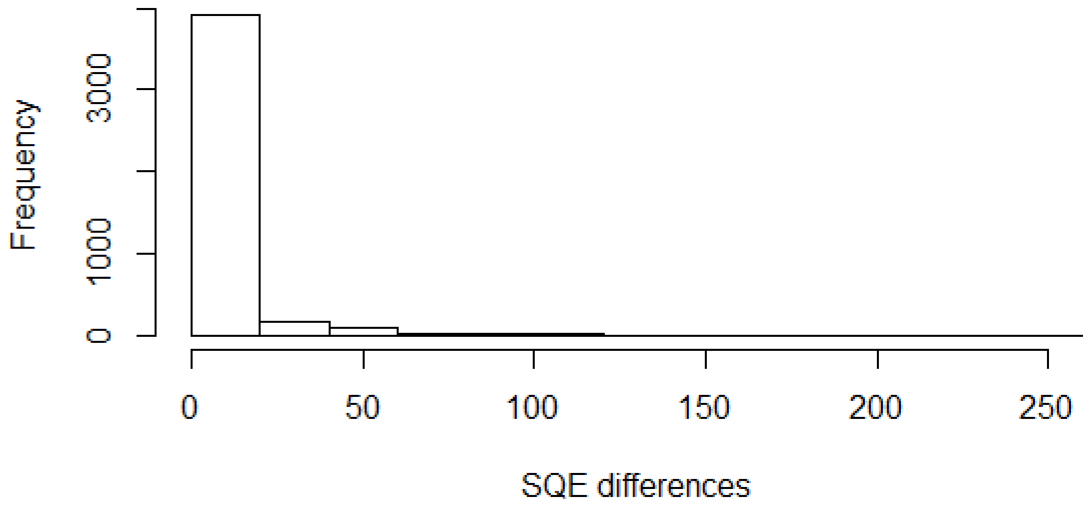
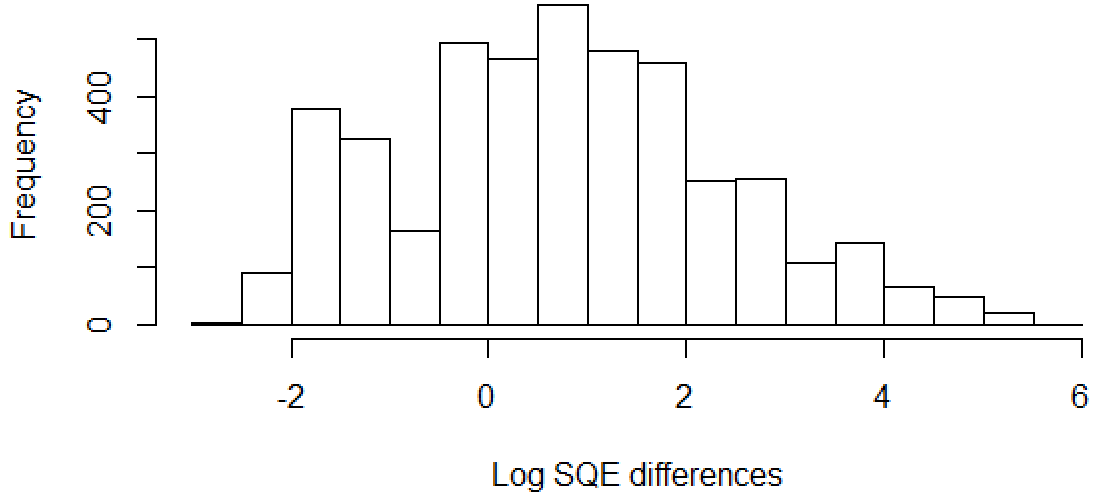


Figure 7: Histogram of log SQE differences



that

the predicted differences must always be positive. Figures 6 and 7 show the untransformed

(Figure 6) and log transformed (Figure 7) data. Using the transformed outcome data instead of the raw data, and refitting model 10, the R^2 improved from .7109 to .8991. At this point the model explains nearly 90% of the variance in the performance of the RMR. A plot of the residuals is in Figure 8. The model is still imperfect, but the plot lacks the very clear groupings of previous plots. Looking at the model parameters, the three-way interaction that had been added previously was no longer significant. Neither was the raw ratio of the standard deviations. The full model is presented in table 9.

A reasonable question after so much exploratory data analysis is "How well does this hold up?" In some pilot data, I had an additional sample of data that replicated many of the conditions (with 1,000 simulations each) in the present analysis and used a different random number seed. Using that replication data with the appropriate log transformation, and the estimated model parameters in Table 9, the R^2 was .9129. The model replicates quite well.

Figure 8: Plot of residual RMR SQE minus the best achievable SQE by condition for model 11

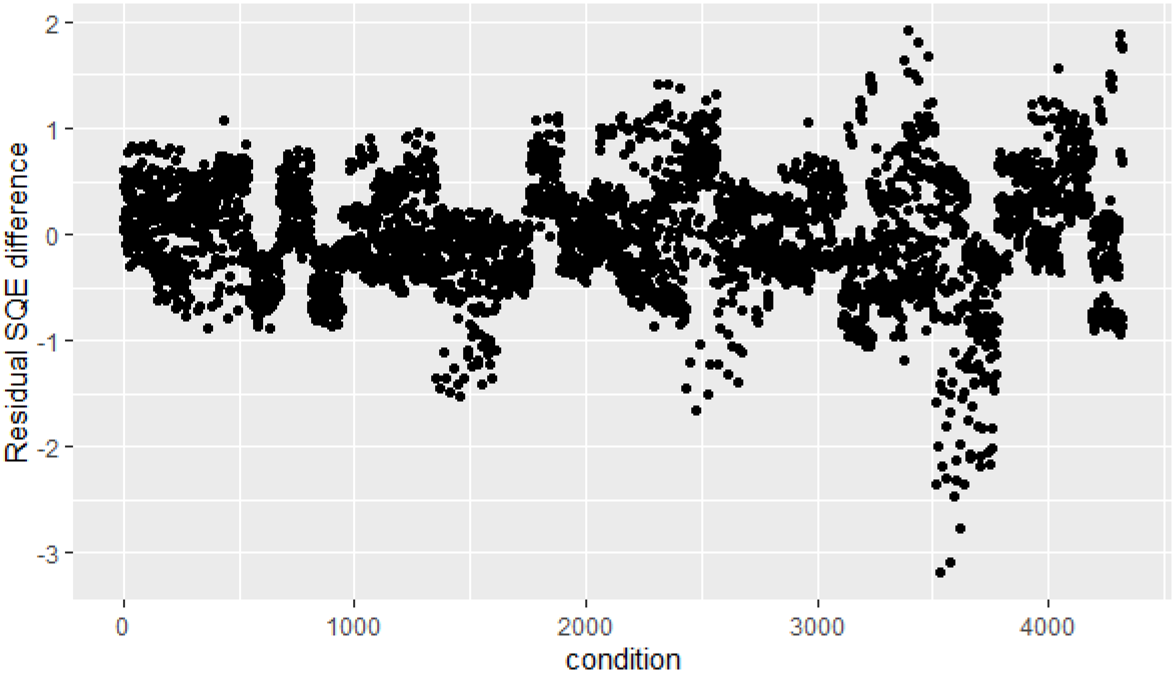
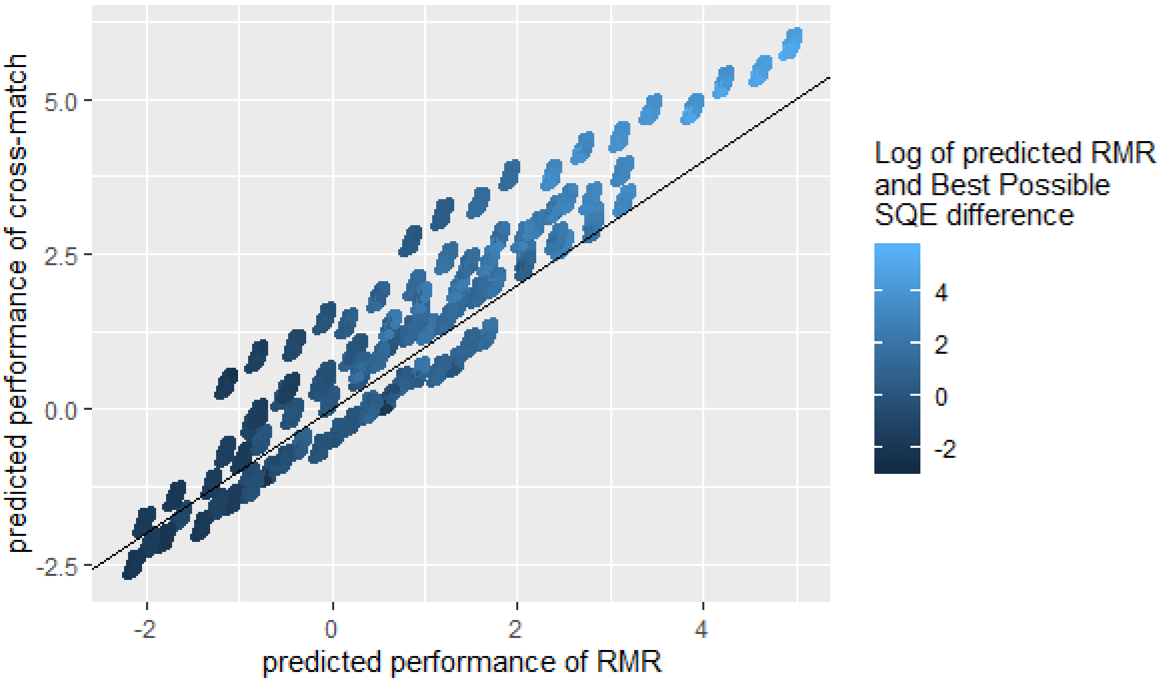


Table 9: Results of final regression model

	Estimate	Standard Error	t value	*p*-value	
(Intercept)	-3.29	0.12	-27.02	< 2e-16	***
Number of variables per factor	1.10	0.05	22.25	< 2e-16	***
Between factor correlations	0.22	0.06	3.50	0.00	***
Baseline Standard deviation	0.27	0.08	3.59	0.00	***
ALRSD	-0.08	0.04	-2.26	0.02	*
Absolute value of Treatment group deviation from baseline within factor correlation	-0.13	0.07	-1.78	0.08	.
Absolute value of the treatment group's within factor correlations	0.13	0.05	2.48	0.01	*
Distribution (1 is log-normal, 0 is normal)	0.33	0.02	21.09	< 2e-16	***
Sample Size	0.00	0.00	-46.17	< 2e-16	***
Raw ratio of group standard deviations	0.00	0.10	-0.05	0.96	
Baseline sd X ALRSD	0.31	0.02	13.80	< 2e-16	***
Baseline sd X Raw sd ratio	0.65	0.06	10.61	< 2e-16	***
Number of variables X Raw sd ratio	0.00	0.04	-0.07	0.94	
Number of variables X baseline sd	0.16	0.03	5.21	0.00	***
Number of variables X baseline sd X Raw sd ratio	0.02	0.03	0.88	0.38	

Conveniently this model consists largely of main effects, which makes interpretation easier. Increasing the number of variables reduces RMR performance, the same is true for increasing the between factor correlations and the baseline standard deviations. ALRSD had a modest, marginally significant, negative effect. A small, marginally significant (*p* = .01) effect for the magnitude of the treated group's within factor correlations was also found. Transitioning from normal to log-normal data hurt RMR performance. Increasing sample size improved model performance (it rounds to 0 in the table, but was -.003). There was a significant positive interaction between the baseline standard deviation and ALRSD, so increasing either would result in increasing decrements to RMR performance. The same was true for the interaction between the number of variables and the baseline standard deviation.

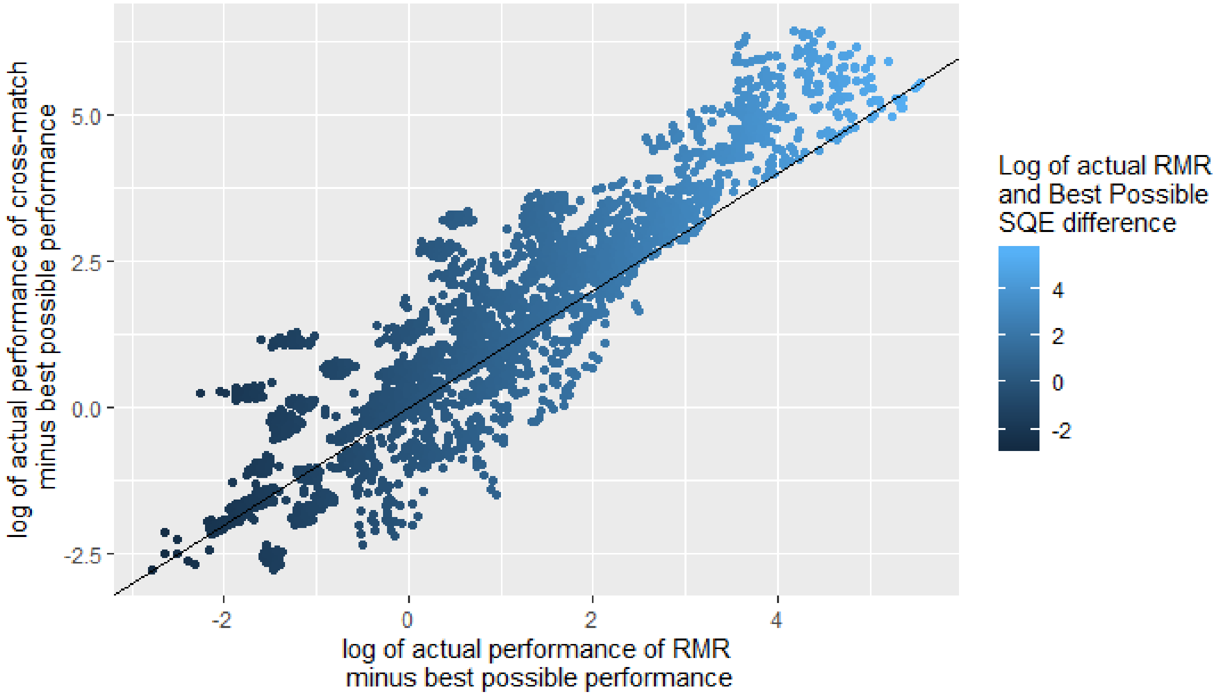
Figure 9: Plot of predicted performance of cross-match vs. RMR



I was interested to know if the same model was equally predictive of other metric's performance, or if the model was substantially different for other metrics. Applying the same model to the SQE difference for the cross-match statistic, but re-estimating the parameters, gave an R^2 of .7496 for the untransformed data, and .9298 for the log transformed data. However, the significance and direction of some of the regression weights was different. I thought it would be useful to plot the predicted values from this model against the predicted values of the RMR.

Figure 9 shows the result. When a point is above the diagonal line the RMR is predicted to have better performance, below the line favors the cross-match statistic. There are few cases (relatively) that favor the cross-match statistic, and those that do are cases where the RMR isn't doing particularly poorly to begin with. Figure 10 shows the result for the actual data. The same result holds. The RMR is usually a better choice and when it is not it still performs adequately. It should be noted that the scales for both plots are log scales. The overall finding points to a possibility for future research. It may be that no one metric strictly dominates other metrics in all cases. Future comparisons of balance metrics, with more conditions than the ones presented here, might be able to provide guidance as to which methods would be optimal under a given data scenario. Researchers would need to limit their advice to predictions that could be made from the raw data prior to estimating any treatment effects.

Figure 10: Plot of actual performance of cross-match vs. RMR



Chapter 4

Discussion

The results suggest a number of useful findings. First, the RMR, despite not having previously been used in propensity score matching, appears to be a highly valuable addition to the current array of balance metrics. This status exists despite the fact that RMR only evaluates second-order moments of the data and takes no account of the mean. It is perhaps surprising that more advanced metrics such as the cross-match statistic did not perform better. I believe that this could be due to a few reasons. The first is that this simulation may simply not have contained cases that played to the cross-match's strengths. The effect of the confounds was limited to main effects, interaction effects and quadratic effects. Other non-linear effects may have been beneficial to the performance of the cross-match statistic.

Second, I hope that this paper provides a framework for evaluating balance metrics in the future. Previous studies have been far too limited in scope. The present study shows that it may be possible to recommend balance metrics on the basis of features of the data, but limited simulations are unable to do this. Even in the current study, a single metric was dominant across most conditions, which makes recommending other metrics difficult.

Third, although it was not a primary research aim of this study, we can compare propensity score matching under a wide variety of conditions and model specifications to randomized trials and unmatched data. While propensity score modeling generally performed better than nothing, none of the propensity score models, even those that ought to have accounted for all data differences, did as well as a randomized controlled trial. Even when limited to cases where the data were normally distributed, the best performing PS model had a mean SQE nearly eight times that of the randomized trial. Further limiting it to cases where the only variables varying between groups were the standard deviation and means of the confounds still resulted in a mean SQE eight times that of a randomized controlled trial. Perhaps a more rigorous matching paradigm (I used 1:1 nearest neighbor matching without a caliper) would improve this performance, but it would necessarily come at a loss of sample size.

I believe future research should focus on expanding our understanding about what conditions can cause a metric to perform better or worse. Specific recommendations are to assess the effect on balance metrics when confounds have variable impact on the outcome of interest. In the current study this weight was equal for all confounds and was one. I also recommend incorporating even higher order moment such as skewness and kurtosis into the assessment of metric performance. I also recommend examining metric performance in the presence of non-continuous variables. The present study used continuous variables throughout, but many variables collected in research are not continuous (e.g., Likert type items, binary response variables, count data, etc.). Another aspect that might be improved is increasing the variability in the performance of the PS models being assessed in a simulation study. In the present study there was variability, but in some conditions it was minimal. In part this limited variability may be due to the fact that in some conditions some terms (e.g., interactions) would have had no effect due to the multivariate structure of the data. For example, in the case where only a single variable was present for each type of confound there were no variables with which they would interact so the correct and incorrect PS model were functionally equivalent.

In conclusion I believe that the RMR is a promising balance metric, despite the lack of previous use. Furthermore, I suspect that other measures of fit from structural equation modeling might find good application in PS modeling. It certainly seems to be the case that PS modelers should stop using univariate measures of balance when easily implemented multivariate alternatives exist. At minimum researchers should consider adding the RMR to statistics they already collect about balance.

Works Cited

Austin, P. C. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in medicine, 33*(6), 1057-1069.

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.

Austin, P. C. (2008). A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine, 27*(12), 2037-2049.

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine*, *26*(4), 734-753.

Barr, D. R., & Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician*, *53*(4), 357-361.

Chen, H., & Small, D. S. (2016). New multivariate tests for assessing covariate balance in matched observational studies. *arXiv preprint arXiv:1609.03686*.

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.

Curran, J. (2018). Hotelling: Hotelling's T^2 Test and Variants. R package version 1.0-5. https://CRAN.R-project.org/package=Hotelling

Dehejia, R. (2005). Practical propensity score matching: a reply to Smith and Todd. *Journal of econometrics*, *125*(1-2), 355-364.

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.

Friedman, J. H., & Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 697-717.

Hansen, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’by Peter Austin, Statistics in Medicine. *Statistics in medicine, 27*(12), 2050-2054.

Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science, 23*(2), 219-236.

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234.

Heller, R., Rosenbaum, P. R., & Small, D. S. (2010). Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, *64*(4), 299-309.

Hill, J. (2008). Discussion of research using propensity-score matching: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003'by Peter Austin, Statistics in Medicine. *Statistics in medicine*, *27*(12), 2055-2061.

Hill, J., Weiss, C., & Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, *46*(3), 477-513.

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*(3), 199-236.

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics, 3, 360--378.* doi:10.1214/aoms/1177732979. https://projecteuclid.org/euclid.aoms/1177732979

Jennrich, R. I. (1970). An asymptotic χ2 test for the equality of two correlation matrices. *Journal of the American Statistical Association*, *65*(330), 904-912.

Kaiser, H. F., & Dickman, K. (1962). Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, *27*(2), 179-182.

Lee, W. S. (2013). Propensity score matching and variations on the balancing test. *Empirical economics*, *44*(1), 47-80.
Maydeu-Olivares, A. & Shi, D. (2019). How to Use the SRMR to Determine the Sample Size Needed for Structural Equation Modeling. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, October, 2019, Baltimore MD.

Nevitt, J., & Hancock, G. R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *The Journal of Experimental Education*, *68*(3), 251-268.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rahmatallah Y., Zybailov B., Emmert-Streib F., & Glazko G (2017). GSAR: Bioconductor package for gene set analysis in R. *BMC Bioinformatics* 18, 61

Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(4), 515-530.

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.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, *79*(387), 516-524.

Sekhon, J. S. (2007). Alternative balance metrics for bias reduction in matching methods for causal inference. *Survey Research Center, University of California, Berkeley*.

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. J Statist. Software 42 (7): 1–52.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Sugiyama, M., Kanamori, T., Suzuki, T., Plessis, M. C. D., Liu, S., & Takeuchi, I. (2013). Density-difference estimation. *Neural Computation*, *25*(10), 2734-2775.