Diagnosing the Diagnostic: A Simulation Study of the Omnibus Test for Covariate Balance

Lauren Vollmer*

Abstract

When selecting a comparison group for an observational study, accurate measures of balance between the treatment group and the matched comparisons are essential. The omnibus test (Hansen and Bowers 2008) complements covariate-specific assessments with a test of simultaneous covariate balance. However, in practice the omnibus test and covariate-specific diagnostics often disagree. Attempting to reconcile these differences, we conduct a simulation study investigating the omnibus test's power and type I error rates under different data-generating scenarios. Across scenarios, the test's power and type I error rates align with theoretical expectations, with one consistent exception: data sets containing predominantly binary variables. We synthesize these findings, including a comparison of the omnibus test and covariate-specific assessments across scenarios, to offer practical guidance for balance diagnostics.

Key Words: causal inference, non-experimental design, propensity score methods

1. Introduction

When estimating causal effects from non-experimental studies, it is typical to develop a comparison group to represent the counterfactual: the outcomes we would have observed from the treatment group in the absence of the intervention. For this counterfactual to hold, the treatment and comparison groups must be well-balanced on the characteristics of interest in the study. It is therefore essential to assess balance rigorously to ensure valid causal inferences.

A rigorous balance assessment examines both univariate balance – whether the treatment and comparison groups are well-matched on each characteristic individually – and multivariate balance – whether the treatment and comparison groups are well-matched across characteristics. Common univariate metrics include standardized differences, which measure the difference in means between the treatment and comparison groups on the standard deviation scale, tests of the hypothesis that the difference in means between the treatment and comparison groups is equal to zero, and tests of the hypothesis that the difference in means between the treatment and comparison groups falls within some practically negligible range. The omnibus test of Hansen and Bowers, which tests whether the difference in means between the treatment and comparison groups falls outside the expected range across all linear combinations of the matching characteristics, is a common multivariate diagnostic.

Univariate and multivariate diagnostics are not mutually exclusive but rather complementary; univariate diagnostics identify possible imbalances on individual characteristics, while multivariate diagnostics guard against scenarios where the treatment and comparison groups are well-matched on each individual characteristic but not across characteristics. However, in practice the two types of diagnostic may not agree; in particular, the omnibus test may reject, implying imbalance,

*Mathematica Policy Research, 955 Massachusetts Avenue, Cambridge, MA 02139

when all standardized differences are quite low, implying the reverse.

At first blush, we may simply dismiss the results of the omnibus test; without any evidence of imbalance on individual characteristics, we cannot pinpoint the variable or variables responsible for the conflicting result. We may also conclude that the omnibus test sets too high a standard by aiming for balance on all linear combinations of the matching characteristics, not all of which, perhaps, are relevant to the problem at hand. Despite these misalignments between the test's design and its practical applications, it serves as a crucial safeguard against pernicious scenarios where bias lurks beneath the surface of an apparently well-balanced sample.

To reconcile these conflicting signals, we must gain a deeper theoretical and practical understanding of the omnibus test. In this paper we aim to address both points, with greater emphasis on the latter. In a first section we will summarize the omnibus test and outline its theoretical underpinnings. In a second section we will develop a simulation study designed to assess the test's empirical properties. In a third section we will summarize the results of the simulation study and in a fourth section we will offer guidelines for practitioners based on these results.

2. The Omnibus Test of Hansen and Bowers

In their 2008 paper, Hansen and Bowers develop an alternative to traditional methods of assessing balance. They begin by examining two traditional approaches, standardized differences and a joint significance test of all relevant characteristics, and finding both wanting. The standardized difference approach produces a list – possibly a long list – of standardized differences and corresponding p-values rather than a single summary statistic. Given that the p-values are likely correlated, it is not immediately clear how to determine whether the sample is balanced based on this information. A joint hypothesis test, by contrast, does produce a single test statistic, but Hansen and Bowers demonstrate by simulation that this statistic is not reliable. In small samples, the joint hypothesis test's Type I error rates are unacceptably high, and Hansen and Bowers conclude after further simulations that deviance tests remain unreliable even with eleven times as many observations as covariates.

As an alternative to these traditional tools, Hansen and Bowers take a fresh look at balance assessments, starting from a simple measure of the difference in means on a single covariate under a single treatment assignment:

$$d_p(\mathbf{z}, \mathbf{x}) = \frac{\mathbf{z}^t\mathbf{x}}{\mathbf{z}^t\mathbf{m}} - \frac{(\mathbf{1}-\mathbf{z})^t\mathbf{x}}{(\mathbf{1}-\mathbf{z})^t\mathbf{m}}$$

In this equation, $\mathbf{z}$ is a vector indicating treatment assignment for each observation and $\mathbf{x}$ is a vector of the same dimension giving the covariate value for each observation. Throughout, Hansen and Bowers emphasize the importance of assessing balance when treatment assignment arises through a clustered design, so they also include $\mathbf{m}$, a vector of cluster sizes.

Although Hansen and Bowers derived the omnibus test to assess balance in a block-randomized design, the notion of clusters is also applicable when evaluating

a matched comparison group for an observational study. The matching process creates sets of one or more treatment units matched to one or more comparison units; as in a block-randomized design, if the matching was successful we can conceive of treatment as randomly assigned within each matched set.

So Hansen and Bowers extend the simple difference-in-means metric $d_p(\mathbf{z}, \mathbf{x})$ to $d(\mathbf{Z}, \mathbf{x})$, which both accounts for clustering and places $d_p(\mathbf{z}, \mathbf{x})$ in the context of its randomization distribution. To conduct such a randomization test, we conceive of $\mathbf{z}$ as a realization drawn from the set of possible treatment assignments $\mathbf{Z}$ and write:

$$d(\mathbf{Z}, \mathbf{x}) = \sum_{b=1}^{B} w_b \left[\frac{\mathbf{Z}_b^t \mathbf{x}_b}{m_{tb}} - \frac{(\mathbf{1} - \mathbf{Z}_b)^t \mathbf{x}_b}{m_b - m_{tb}} \right]$$

Here we introduce additional notation: $b \in \{1, 2, \ldots, B\}$ denotes a block, or matched set in the observational case; m_{tb} represents the expected number of treated units in cluster b, while $m_b - m_{tb}$ represents the number of untreated units in the same cluster. Likewise, $\mathbf{x}_b$ represents the vector of x-values for the units in block b.

Hansen and Bowers note that under the relevant central limit theorems $d(\mathbf{Z}, \mathbf{x})$ converges to a Normal distribution in large samples. Extending this univariate statistic to the multivariate case, then, we have

$$d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k) := [d(\mathbf{z}, \mathbf{x}_1), \ldots, d(\mathbf{z}, \mathbf{x}_k)] \cdot \left\{ Cov \quad \begin{bmatrix} d(\mathbf{Z}, \mathbf{x}_1) \\ \ldots \\ d(\mathbf{Z}, \mathbf{x}_k) \end{bmatrix} \right\}^{-1} \begin{bmatrix} d(\mathbf{z}, \mathbf{x}_1) \\ \ldots \\ d(\mathbf{z}, \mathbf{x}_k) \end{bmatrix}$$

As desired, this test produces a single summary statistic and corresponding p-value that describe the difference in means across all linear combinations of the characteristics included in the balance assessment. Because the univariate components $d(\mathbf{Z}, \mathbf{x}_1), \ldots, d(\mathbf{Z}, \mathbf{x}_k) \sim$ Normal, $d^2(\mathbf{z}; \mathbf{x}_1, \ldots, \mathbf{x}_k) \sim \chi^2$, where the degrees of freedom are equal to the rank of the variance-covariance matrix. This approximate distribution facilitates derivation of p-values and thus rejection or non-rejection of the hypothesis that the difference in means across all linear combinations of characteristics falls within the expected range.

Based on the properties of this test, Hansen and Bowers assert that the χ^2 approximation is relatively reliable even in small samples. They also claim that the test is robust to expansion of the set of characteristics to examine; assessing balance on additional characteristics will not lead the test falsely to reject the null hypothesis of good balance.

3. A Simulation Study

Hansen and Bowers advocate for their test on the theoretical grounds described above, under the assumption that the characteristics $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are normally distributed. In practice, however, the data used to select a comparison group, and thus used to assess post-matching balance, may deviate substantially from the test's theoretical expectations. In fact, as already noted, in practice the omnibus test often

contradicts the findings of other univariate diagnostic measures like standardized differences; uniformly low standardized differences imply that balance is good across all covariates, while a low omnibus test p-value implies the opposite.

In this study we seek to investigate the performance of the omnibus test empirically, determining whether such contradictions are more likely to arise in certain scenarios such as data sets with many observations, data sets with many characteristics, or data sets that violate the test's assumptions to varying degrees.

3.1 Simulation Parameters

We identified four primary data characteristics that could plausibly contribute to the performance of the omnibus test:

- Number of observations $n \in \{100, 250, 500, 1000, 10000\}$
- Number of variables $p \in \{10, 25, 50, 100\}$
- True difference in means $d \in \{0, 0.5, 1, 2\}$ on the standard deviation scale, constant across variables
- Distributional composition

The first three simulation parameters, n, p, and s, are straight-forward; the latter less so. The fourth parameter governs the probability distributions used to generate the simulated data, which we seek to vary to determine whether, for example, skewed or binary data, which violate the test's normality assumption but which we often observe in practice, alter the test's behavior.

To assess the test's vulnerability to non-normal data, we constructed seven types of data sets based on the proportions of variables generated according to three distributions: normal, representing the expected data type; gamma, representing highly skewed data that is ubiquitous in financial, health care, and other applications; and binary, representing crucial background characteristics like gender. We present the seven distributional types in Table 1.

Table 1: Distributional Composition of Simulated Data Sets

Type	Normal	Gamma	Binary
A	20%	0%	80%
B	20%	40%	40%
C	20%	80%	0%
D	60%	0%	40%
E	60%	20%	20%
F	60%	40%	0%
G	100 %	0%	0%

In data sets of type A, for example, 20% of the variables are generated according to a normal distribution and 80% of the variables are generated from a binary distribution.

3.2 Simulation Structure and Execution

Each combination of these simulation parameters defines a simulation scenario. To ensure stability in the simulated data set, we exclude scenarios where the number of observations is less than ten times the number of variables, i.e. $n < 10p$, which gives 392 distinct scenarios. We perform 1,000 iterations for each simulation scenario; for each iteration, we record both the omnibus test p-value and the standardized difference in means for each variable in the simulated data set.

In each iteration of the simulation, we simulate a data set according to the parameters, drawing n observations to represent each of p variables, proportionally from the distributions determined by the distributional composition parameter. For example, if $n = 1000$, $p = 25$, and the distributional type is A, we create 5 normally distributed variables and 20 binary variables, each with 1000 observations.

If the true difference in means d in the given iteration is zero, we generate a single data set with n observations and p variables, each of which has mean zero. We then randomly assign a binary treatment vector to classify observations as treatments or comparisons and perform the omnibus test comparing variable means by treatment assignment. To produce the desired number of iterations, we simply permute the treatment assignment 1,000 times and re-calculate the omnibus test p-value and standardized differences for each fresh permutation.

If the true difference in means $d > 0$, we generate $\frac{n}{2}$ observations for each variable in the manner described above and assign these observations to the treatment group, then generate the remaining $\frac{n}{2}$ observations assigned to the comparison group from distributions with a mean of d standard deviations, where standard deviation is defined according to the distribution of the variable. We calculate the omnibus test p-value and variable-specific standardized differences based on this treatment assignment. Unlike in the $d = 0$ case, however, we cannot simply permute the treatment assignment lest we randomly create a data set with $d = 0$. We must simulate a new data set for each iteration when $d > 0$.

We coded and executed the simulation in R, using Hansen and Bowers' package {RItools} to calculate both the omnibus test and standardized differences.

4. Results

When examining the results of the simulation, we focused on two general areas: its overall properties, such as Type I error and power, and scenarios in which contradictions with univariate balance diagnostics were most likely.

4.1 Overall Properties

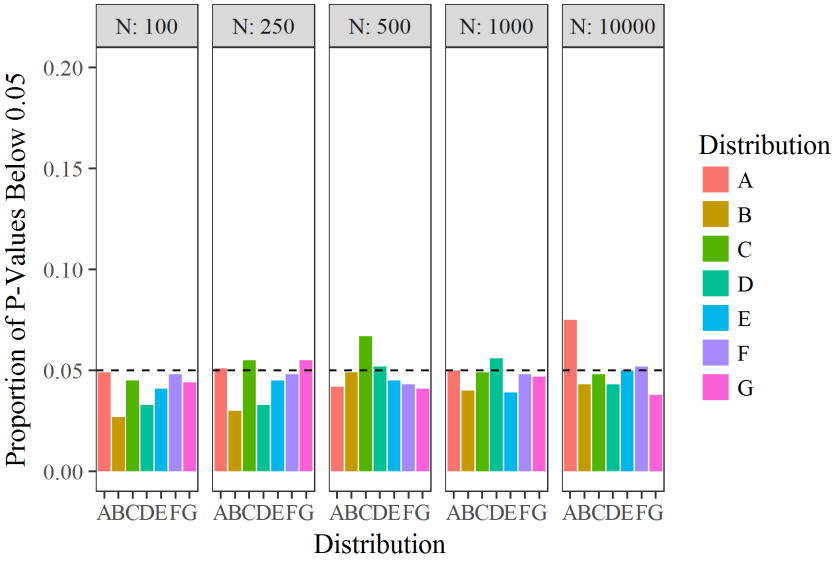
4.1.1 Type I Error

We assessed the omnibus test's Type I error rate by inspecting the results of the simulations in which the true difference in means between the treatment and comparison groups was set to 0. In these simulations, any omnibus test p-value < 0.05 represents a Type I error.

This question is of particular interest because detractors of the test commonly cite its over-sensitivity to large sample sizes. In the popular conception, the omnibus test is more likely to reject when the sample is large, even if the differences in means across variables are negligible. Thus, we would expect to see a higher Type I error rate in larger samples than in smaller ones.

However, as we see in Figure 1, the Type I error rate is quite consistent across scenarios, with no evident patterns based on sample size or distributional composition.

Figure 1: Type I Error Rate by Sample Size and Distributional Composition



In this figure, each panel represents a given sample size, ranging from 100 on the left to 10000 on the right. In each panel, the x-axis specifies the distributional composition of the data and the y-axis represents the proportion of p-values below 0.05. The dashed horizontal line marks the expected Type I error rate of 0.05, and for all sample sizes, the Type I error rates largely fall at or below this line. Counter to expectation, then, the omnibus test has the expected Type I error rate at all sample sizes investigated.

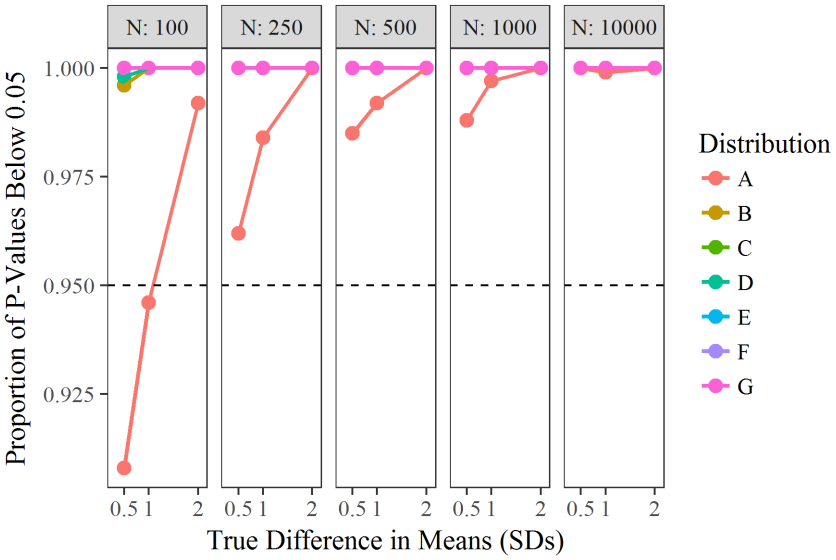
4.1.2 Power

Conversely, we might wish to learn about the test's power in different simulation scenarios. If the test is overpowered at large sample sizes, we might observe the behavior described in which the test rejects the null hypothesis of good balance even when the differences in means are minuscule.

We assessed the test's power by calculating the proportion of iterations in which the omnibus test p-value is less than 0.05 in simulation scenarios where the true difference in means is greater than zero. In Figure 2 we plot the probability of correctly rejecting the null for each true difference in means, sample size, and distributional composition.

As in Figure 1, each panel in this figure represents a different sample size, from

Figure 2: Power by True Difference in Means, Sample Size, and Distributional Composition



lowest on the left to highest on the right. In each panel, the x-axis represents the true difference in means on the standard deviation scale, while the y-axis represents the proportion of iterations in which we observed a p-value less than 0.05. Different-colored lines represent data sets of different distributional types. The dashed horizontal line at 95% serves as a reference.

Power is extremely high across all sample sizes and is quite consistent across data sets of different types. In the first panel, corresponding to data sets with only 100 observations, we see slightly lower power for data sets of types B (20% normal, 40% Gamma, 40% binary) and D (60% normal, 40% binary) when the true difference in means is only 0.5 standard deviations, but in the remaining panels the magnitude of the difference in means does not affect the test's power.

The orange lines, which track the power for data sets of distributional type A (20% normal, 80% binary), diverge markedly from this pattern. Although power is quite high – over 90% – for all combinations of sample size and true difference in means, it increases sharply as the magnitude of the difference in means increases. This steep slope is most evident in simulation scenarios with comparatively small data sets and attenuates as the sample size increases. For small data sets with many binary variables, then, the omnibus test less reliably detects small differences in means.

4.2 Contradictions between Omnibus Test and Standardized Differences

Understanding the general properties of the omnibus test deepens our appreciation of its reliability without necessarily offering insight into specific practical situations. In particular, in practice we often observe contradictions between the omnibus test and standardized differences or other univariate measures of balance. We seek to reconcile those contradictions by probing more deeply into the simulation scenarios

most likely to produce them.

To determine what scenarios are most likely to result in contradictions between the omnibus test and standardized differences, we first sought to summarize standardized differences across the variables assessed. We applied two criteria:

- **Relaxed Criterion**: at least 85% of standardized differences are < 0.25 and at least 60% are < 0.1 in absolute value
- **Strict Criterion**: at least 90% of standardized differences are < 0.25 and at least 70% are < 0.1 in absolute value

These criteria are necessarily somewhat arbitrary, since the thresholds determining adequate balance will vary based on the application. However, standardized differences of less than 0.25 are generally understood to represent acceptable balance (see, for example, Stuart 2010), while standardized differences of less than 0.10 represent good balance.

We wish to determine whether any simulation iterations in which the omnibus test rejects the null hypothesis – that is, in which the omnibus test indicates imbalance – meet either of these two criteria. We tabulate the scenarios in Table 2.

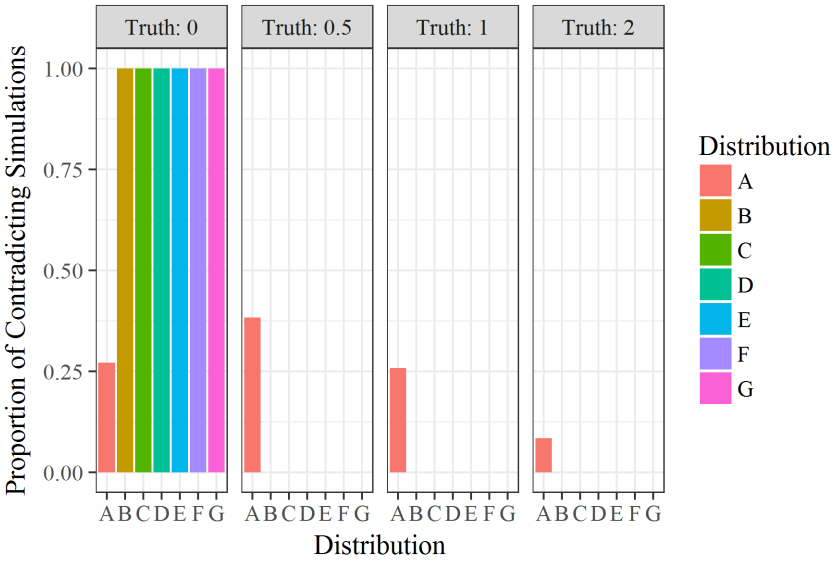
Table 2: Simulations by Omnibus Test Significance and Standardized Difference Criteria

Omnibus test	Criterion Met	Criterion Not Met
Relaxed Criterion		
Does not reject	86%	14%
Rejects	**2**%	98%
Strict Criterion		
Does not reject	76%	24%
Rejects	**1**%	99%

As Table 2 shows, in the vast majority of cases the omnibus test and standardized differences reach the same conclusion. Nonetheless, balance meets the relaxed criterion in about 2 percent of cases where the omnibus test rejects, while balance meets the strict criterion in about 1 percent of cases where the omnibus test rejects. We seek to identify any patterns that might aid the practitioner in deciphering these contradictions when they arise in practice.

In Figure 3 we plot the proportion of contradictory cases based on the strict criterion by distributional composition and true difference in means; although we do not show the corresponding plot based on the relaxed criterion, the over-arching trend is the same in both cases.

Figure 3: Proportion of Contradictory Simulations by True Difference in Means and Distributional Composition, Strict Criterion

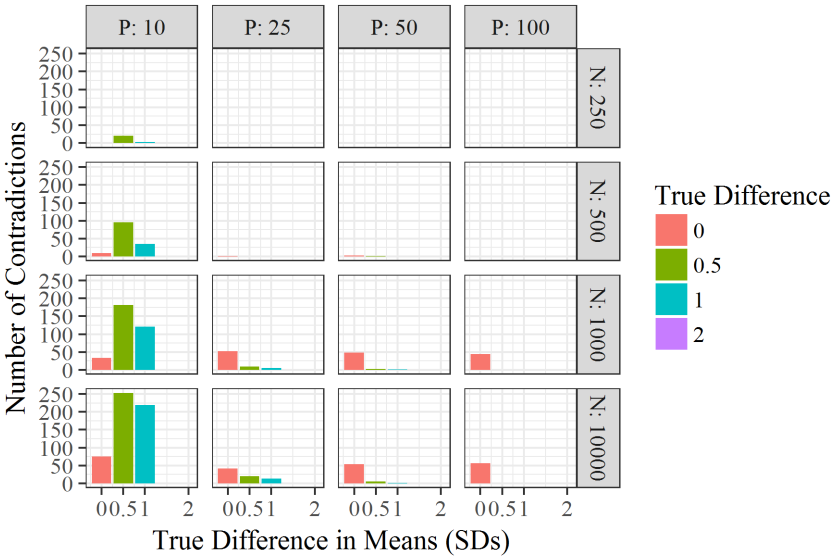


In this figure, each panel represents a different true difference in means, from zero – scenarios where the standardized difference criterion has successfully identified good balance – to two standard deviations. On the y-axis of each panel we show the proportion of contradictory simulations of that type; the x-axis gives the distributional type of the data. For most data set types, 100% of contradictory simulations appear in the left-most panel, where the true difference in means is equal to zero. In these cases, the contradiction arises through a Type I error, so the interpretation based on standardized differences is correct.

These results, then, seem to bear out the conventional conception that standardized differences are a more reliable barometer of balance than the omnibus test. For distributions of type A – where 80% of the variables are binary – we reach exactly the opposite conclusion; in these cases, contradictions between the omnibus test and the standardized difference criteria are more likely when the true difference in means is *greater* than zero. Thus, for data sets with primarily binary variables, the omnibus test may be more a more reliable balance diagnostic than standardized differences.

The pattern of results in Figure 3 suggests that for data sets with primarily binary variables, these contradictions are more likely when the true difference in means is relatively small – on the order of 0.5 standard deviations. Other factors, like the sample size or the number of variables in the balance assessment, may also contribute to the observed patterns. We plot the number of contradictory simulations for data sets of type A by each of these factors in Figure 4.

Figure 4: Breakdown of Contradictory Simulations of Distributional Type A



In this figure, the columns represent the number of variables in the data set, ranging from 10 on the left to 100 on the right, while the rows represent the number of observations in the data set, from 250 on the top to 10000 on the bottom. The y-axis gives the number of simulations and the x-axis gives the true difference in means. We observe immediately that the contradictory simulations are concentrated in the first column, where the number of variables in the balance assessment is quite low, and toward the bottom, where the number of observations is large.

As the number of variables increases, Type I errors – cases where the true difference in means is zero – predominate. We note that we do not see evidence of contradictory simulations when the sample size is small, perhaps because, as already discussed, the omnibus test is not as well-powered for smaller data sets containing primarily binary variables.

Even for data sets containing primarily binary variables, contradictions between the omnibus test and standardized differences are most likely to be Type I errors if the balance assessment covers a moderate to large number of variables. When the balance assessment covers a large data set with comparatively few variables, however, the reverse is true: contradictions between the omnibus test and standardized differences are more likely to denote true imbalances.

5. Discussion

Through this simulation study, we have confirmed that the omnibus test's theoretical claims hold in practice: the test has the expected Type I error rate of 5% regardless of sample size and boasts power well above 90% across data sets of different sizes and distributional types. In fact, lending credence to popular conception, the test may be overpowered for large sample sizes; power approaches 100% even for comparatively small differences in means when the sample size is large. Further work could investigate the smallest detectable difference in treatment and comparison group means at different sample sizes to aid practitioners in calibrating their interpretation of omnibus test results.

These simulations also shed further light on scenarios in which the omnibus test contradicts the conclusions of other univariate balance metrics. In particular, we are interested in scenarios where univariate measures, like standardized differences, imply good balance across all the relevant variables, but the omnibus test rejects. Our simulations show, first, that these contradictions are very rare – no more than 2 percent of cases in which the omnibus test rejects. Second, for most data set types, these contradictions arise through Type I errors, so standardized differences provide a more reliable indication of the quality of balance in the matched sample.

Data sets comprising primarily binary variables are the exception to this rule. For such data sets, the omnibus test and standardized differences are most likely to conflict when the true difference in means is greater than zero, so the omnibus test most accurately describes the balance in the sample. These misleading standardized differences are most likely to occur when the data set is large and when the balance assessment covers relatively few variables. When the balance assessment covers 25 or more variables, disagreement between the two diagnostics is most likely to be a Type I error.

5.1 Limitations and Extensions

In these simulations, we made several simplifying assumptions that limit the extent to which the simulated data sets mimic data practitioners are likely to encounter. First, we generated each variable in the data set independently at random, while in reality variables are likely to be correlated – in a health care setting, for example, we would expect to see relationships between a patient's age and medical risk. Second, the true difference in treatment and comparison group means was constant across all variables in each simulated data set. In practice, matching may achieve better balance on some variables than on others. Finally, though the omnibus test's ability to assess balance on all linear combinations of variables recommends it as a crucial supplement to univariate diagnostics, we induced imbalance within rather than across variables.

Extensions of this work should investigate more complex scenarios allowing for varying degrees of correlation among variables and varying degrees of balance on different subsets of variables. Crucially, future work should also examine scenarios in which matching achieves balance within each variable but not across variables, since it is in these cases that the omnibus test earns its keep.

5.2 Conclusions: Guidance for Practitioners

These results point to guidelines for statisticians and policymakers interpreting balance diagnostics. The omnibus test's empirical properties are very strong, recommending it as a supplement to conventional univariate diagnostic measures. In situations where these two types of diagnostics contradict each other, Type I error is the most likely explanation, so univariate measures should take precedence unless most of the variables included in the balance assessment are binary. For data sets with comparatively few, predominantly binary variables, contradictions between the omnibus test and standardized differences are more likely to occur when the treatment and comparison groups are in fact imbalanced. In these cases, the omnibus

test is more likely to depict the sample balance accurately.

These guidelines do not, however, address some of the fundamental practical concerns with the omnibus test. When evaluating balance based on standardized differences, practitioners and subject-matter experts can apply their judgment and expertise in determining what variables are most important and what magnitude of difference is meaningful for each variable. The omnibus test, which weights each variable and each linear combination of variables equally, lacks this flexibility. Marrying the theoretical and empirical strengths of the omnibus test with the flexibility of the univariate diagnostics, a weighted version of the omnibus test, through which users could prioritize individual variables and their linear combinations, would represent a substantial addition to the diagnostic toolbox.

REFERENCES

Hansen, B. B., and Bowers, J. (2008), "Covariate Balance in Simple, Stratified and Clustered Comparative Studies," *Statistical Science 23*, 2, 219–236.

Stuart, E.A. (2010), "Matching methods for causal inference: a review and a look forward," *Statistical Science 25*, 1, 1–21.