Winsorizing and Trimming with Subgroups

Till Wicker*
Tilburg University, Warandelaan 2,
5037 AB Tilburg, The Netherlands

September 29, 2025

*Corresponding Author. I am grateful to Giuseppe Musillo, Anaya Dam, Juan Segnana, Hazal Sezer, Manon Delvaux, Christoph Walsh, Ashley Wong, Daan van Soest, Patricio Dalton, and David McKenzie for their helpful comments and suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

Winsorizing and trimming are used to minimize the effects of outliers on estimated treatment effects. The typical approach winsorizes/trims the tails of the whole sample, even if there are heterogeneous subgroups within the sample -– like a treatment and control group in Randomized Controlled Trials. An alternative approach – *Stratified Winsorizing/Trimming* – winsorizes subgroups separately, ensuring that an equal proportion of observations are winsorized/trimmed per subgroup. Monte Carlo simulations of an RCT illustrate that *Stratified Winsorizing/Trimming* reduces the treatment effect bias and risk of Type II errors compared to the traditional approach, although at the cost of a greater likelihood of Type I errors. Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the chosen winsorizing/trimming technique can affect the magnitude and statistical significance of treatment effects. Practical guidelines for researchers wanting to winsorize/trim a sample that consists of heterogeneous subgroups are discussed.

Key words: Winsorizing, Trimming, Biased Treatment Effect, Type I Errors, Type II Errors.
JEL codes: C18, C21, C81

1 Introduction

Researchers are concerned with the role of measurement errors and outliers in the estimation of variables and treatment effects. For example, Gollin and Udry (2021) find that measurement errors and productivity shocks explain between half and two-thirds of the variance in productivity among farmers in Uganda and Tanzania. While one literature strand focuses on designing surveys to minimize the occurrence of measurement errors, another strand focuses on dealing with measurement errors — in particular, outliers — once the data is collected.[1] The most common approach to mitigating the role of outliers is to winsorize or trim the tails of the sample distribution. Winsorizing entails "replacing any values bigger than a certain percentile with the value of the data point at that percentile itself", while trimming consists of "replacing the outliers with a missing value" (World Bank, 2023).[2]

Most researchers winsorize/trim the whole sample, but some recent papers – including Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023) – winsorize/trim subgroups separately, for example by winsorizing/trimming treatment and control groups of a Randomized Controlled Trial (RCT) individually. This paper explores the advantages and disadvantages of winsorizing/trimming the whole sample versus separate subgroups (called *Stratified Winsorizing/Trimming*). After outlining the two techniques in Section 2, Monte Carlo simulations of an RCT in Section 3 illustrate the effects of both winsorizing/trimming techniques on a study's estimated treatment effect bias and the likelihood of Type I and II errors.[3] The simulations reveal that compared to the standard approach of winsorizing/trimming the whole sample, *Stratified Winsorizing/Trimming* increases the likelihood of Type I errors, while reducing both the bias on the treatment effect estimate and the likelihood of Type II errors. The two approaches to winsorizing/trimming are subsequently applied to Angelucci et al. (2023) and Jack et al. (2023) in Section 4 to illustrate that the chosen winsorizing/trimming method can impact both the magnitude and statistical significance of estimated treatment effects in RCTs as well as

[1] For example, the Journal of Development Economics released a Special Issue on Measurement and Survey Design.

[2] Other terminology used includes truncating (both for winsorizing and trimming), and replacing data with empty observations (for trimming).

[3] The focus of the simulations is on winsorizing, as this is more commonly applied in the academic literature. However, the same intuition and results hold for trimming, see Appendix A.

Difference-in-Difference designs. Section 5 discusses practical guidelines associated with winsorizing/trimming the whole sample versus separate subgroups, including Stata and R code, before Section 6 concludes.[4]

By focusing on the most common method of dealing with outliers, this paper contributes to the literature on the importance of outliers and measurement errors in the estimation of variables and their relationships. While quantile treatment effects are often used to highlight the heterogeneity of treatment effects across a sample distribution, trimming and winsorizing are used to reduce the effects of outliers. For example, Angrist and Krueger (2000) apply trimming to matched employer-employee data and conclude that "a small amount of trimming could be beneficial" to reduce the effect of outliers. Bollinger and Chandra (2005) illustrate that winsorizing and trimming can result in biased regression estimates, by inducing a sample selection bias: the remaining sample post-winsorizing/trimming is no longer representative of the underlying population (Heckman, 1979; Goldberger, 1981; Heckman, 1990). This paper contributes to this literature by identifying an additional potential bias with the traditional approach to winsorizing/trimming the whole sample as a result of the unequal winsorizing/trimming of subgroups of the sample, and illustrates the advantages and disadvantages of both winsorizing/trimming techniques on biased estimates of treatment effects, and the likelihood of Type I and II errors.

More recently, Broderick et al. (2023) and Young (2019) have placed renewed emphasis on how outliers and *high leverage* observations can affect average treatment effects. Broderick et al. (2023) show that dropping less than 1% of observations can change the magnitude and sign of estimated treatment effects of published economics papers. Young (2019) illustrates that, across 53 papers published in AEA journals, removing just a single observation results in 35% of treatment effects that were statistically significant at the 1% level to no longer be as statistically significant. This paper contributes to the literature on the sensitivity of treatment effect estimates to outliers by illustrating how the winsorized/trimmed outliers can affect the treatment effect estimate, with empirical applications to Angelucci et al. (2023) and Jack et al. (2023). Across both papers, treatment effect estimates change by 53.84% on average as a result of *Stratified Winsorizing/Trimming* instead of the

[4]The Online Appendix reproduces Monte Carlo simulations for trimming, a theoretical framework, and applications of both winsorizing techniques to Schilbach (2019) and Augsburg et al. (2015).

traditional approach of winsorizing/trimming the whole sample. Reporting treatment effects as a result of both winsorizing/trimming techniques can complement the "Approximate Maximum Influence Perturbation" of Broderick et al. (2023) to strengthen the robustness of treatment effect estimates.

Based on the Monte Carlo simulations and applications to Angelucci et al. (2023) and Jack et al. (2023), this paper offers six practical guidelines for researchers who want to winsorize/trim outliers, further outlined in Section 5:

1. Irrespective of the empirical strategy, panel data collected during different time periods/survey rounds should be treated as separate subgroups, and hence winsorized/trimmed separately.

2. With Randomized Controlled Trials, there is no clear winner between winsorizing/trimming the entire sample vs. stratifying per subgroup. Instead, reporting both techniques provides a more robust estimation of the treatment effect.

3. For Difference-in-Difference and Regression Discontinuity Designs, the recommendation is to use *Stratified Winsorizing/Trimming* as the study sample consists of different subgroups.

4. Reporting the proportion of winsorized/trimmed observations per subgroup in a paper's appendix can alleviate concerns that observations in certain subgroups are disproportionately winsorized/trimmed.

5. For Pre-Analysis Plans of RCTs, the recommendation is to pre-specify that both approaches to winsorizing/trimming will be used as a pre-specified percentile cut-off, in order to provide further robustness that treatment effect estimates are not driven by outliers.

6. Subgroups should be categorized by time periods (in the case of panel data), and "treatment" groups. The only exception is the baseline of an RCT, where it is known that the treatment and control groups are drawn from the same underlying distribution.

2 Winsorizing and Trimming: The Basics

Outliers, particularly in self-reported data, can arise for a variety of reasons: enumerator fatigue, human error, or misreporting, to name a few. Regardless of their reason, outliers can result in the sample distribution differing from the true, unobserved population distribution. Similarly, outliers – in particular *high leverage observations* (Broderick et al., 2023) – can bias treatment effect estimates. Therefore, authors frequently winsorize/trim outliers (the shaded region in Figure 1) such that the observed sample distribution more closely reflects the true, unobserved population distribution.

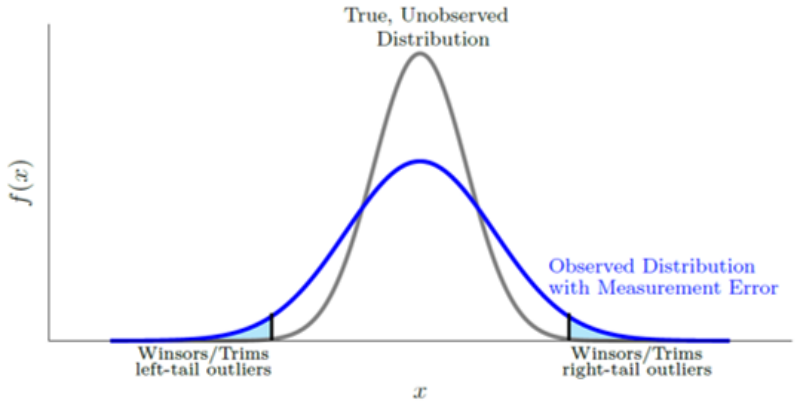
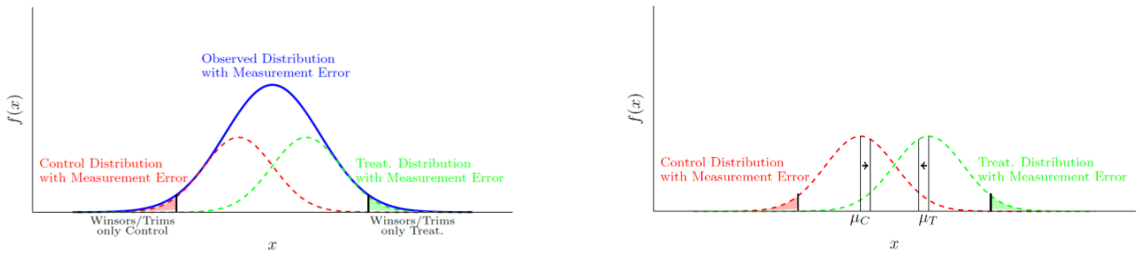


Figure 1. Winsorizing/Trimming the Whole Sample

The most common approach to winsorizing/trimming is to define an upper and/or lower percentile bound beyond which observations are considered outliers and hence winsorized/trimmed. However, some studies use different criteria for winsorizing/trimming their data, informed by the underlying data generating process. For example, Allcott et al. (2020) winsorize individual's willingness-to-accept to abstain from Facebook at \$170, as that was the upper bound of the distribution of Becker–DeGroot–Marschak offers made. de Mel et al. (2019) trim a firm's number of workers at 5, in order to be powered to detect small changes in the outcome variable, due to a long right tail. Fafchamps et al. (2012) trim observations above 10,000 Ghanaian cedi, arguing these are likely due to currency errors. For situations like these, a clear rationale exists to winsorize/trim at a certain value. However, often outcome variables are winsorized/trimmed at the 95th or 99th percentile to account for right-tailed outliers, without an understanding of the data generating process

and cause of the outliers. Particularly with the emergence of Pre-Analysis Plans, researchers pre-specify how they will deal with outliers, without understanding the underlying nature of these outliers, and hence rely on rules of thumb.[5]

The traditional approach to winsorizing and trimming treats the sample as one distribution, even when the sample consists of subgroups, such as a control and treatment group in an RCT.[6] If the measurement error is uncorrelated with the subgroup (e.g., the result of an enumerator error, or white noise) – as is typically the case – when authors winsorize/trim, the expectation is that the likelihood of outliers and measurement errors is the same across subgroups within the sample. However, if the subgroups have different distributions – for example due to a non-zero treatment effect – winsorizing or trimming the whole sample can disproportionately winsor/trim the tails of the distribution of each subgroup. Figure 2a illustrates this in the case of an RCT where the treatment group experiences a positive treatment effect, where the traditional approach to winsorizing/trimming trims the bottom-tail of the control group distribution, and the upper-tail of the treatment group distribution.[7] If the measurement error is uncorrelated with the subgroup (e.g., the result of an enumerator error, or white noise) – as is typically the case – the differential trimming of outliers in the treatment and control distributions can generate a biased treatment effect, as illustrated by Figure 2b.



(a) Treatment vs. Control **(b)** Treatment Effect Bias

Figure 2. Winsorizing/Trimming: by subgroup

[5] As of February 21th, 2025, 32% of the Pre-Analysis Plans Accepted during a Stage 1 Review at the Journal of Development Economics specified that they intend to winsorize or trim the data.

[6] Other examples include those above vs. below the cutoff in a RDD design, those receiving an intervention vs. not in a DiD design, data points collected at different time intervals, heterogeneity by gender/race, etc.

[7] Alternatively, Figure 2a can also illustrate the case of Wave I vs. Wave II of a survey. Similarly, Figure 2a could also represent underlying differences between two groups in a Difference-in-Differences empirical strategy that nevertheless satisfy the parallel trends assumption.

In the stylistic example of Figure 2a, winsorizing/trimming the left and right tail of the sample distribution results in winsorizing/trimming the left tail of the control group distribution, and the right tail of the treatment group distribution. The implications of the differential winsorizing/trimming of subgroups is illustrated in Figure 2b, which shows that the means of both subgroups move inwards. This can result in a biased underestimation of the true treatment effect.

An alternative winsorizing/trimming technique – *Stratified Winsorizing/Trimming* – instead winsorizes/trims each subgroup separately, as illustrated in Figure 3. By ensuring that an equal proportion of observations are winsorized/trimmed from each subgroup (and an equal proportion of left- and right-tailed observations per subgroup), the distribution of each subgroup more closely reflects the underlying population distribution of these subgroups (see Figure 3).

The next section illustrates, using Monte Carlo simulations, the effects of winsorizing the whole sample vs. subgroups separately on treatment effect estimate biases, a study's statistical power (Type II errors), and Type I errors.

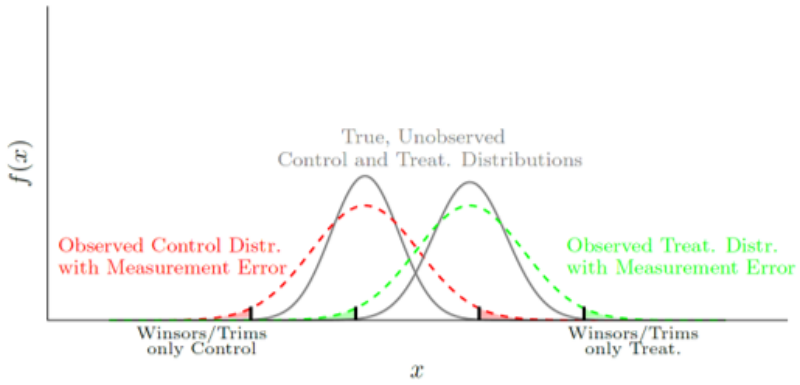


Figure 3. Stratified Winsorizing/Trimming

3 Monte Carlo Simulations

Monte Carlo simulations replicate an RCT where 500 participants are randomly assigned to a control and a treatment group.[8] The estimated regression is $Y_i =$

[8]Monte Carlo simulations simulate a RCT, as this is the most common empirical method where winsorizing/trimming is used. Furthermore, the major limitation of *Stratified Winsorizing/Trimming* – the increased likelihood of Type I errors when the stratified groups are actually from the same underlying distribution – does not apply to other empirical strategies such as DiD or RDD.

$\alpha + \beta_1 T_i + \varepsilon_i$, where T_i is an indicator equal to one if the participant is assigned to the treatment group, and zero otherwise. β_1 therefore is an unbiased estimate of the treatment effect. The error term is standard normally distributed ($\sim N(0,1)$), while the outcome variable Y_i is winsorized at the 90% level (top and bottom 5%), using the traditional approach of winsorizing the whole sample, as well as *Stratified Winsorizing* separately by treatment group.[9] Due to the nature of the simulations, outliers are uncorrelated with assignment to the treatment or control group.

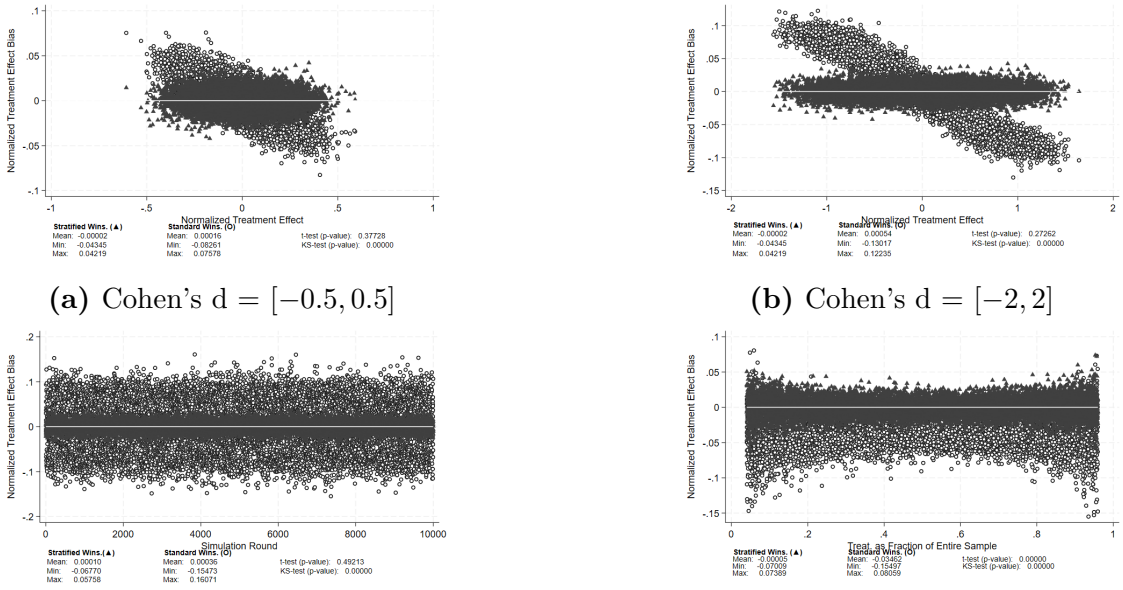
3.1 Biased Treatment Effects

The stylistic example of Figure 3 illustrates how winsorizing the entire sample distribution can differentially trim subgroups if their underlying distributions differ. This in turn can bias the treatment effect estimate by under-reporting the true treatment effect. To test this, I run 10,000 simulations of the RCT with 500 subjects divided across a treatment and control group. Each simulation generates a treatment effect estimate (β_1) without winsorizing, and the two approaches to winsorizing. The resulting bias is measured as the difference in treatment effects (between the non-winsorized sample, and the winsorized sample, done separately for the two approaches to winsorizing), normalized by the standard deviation of the control group of the non-winsorized sample. The horizontal white line means there is no treatment effect bias as a result of winsorizing. Values above the white horizontal line indicate that winsorizing induces a positive bias on the treatment effect estimate, while values below the horizontal line indicate a negative bias. Results are presented in Figure 4.

Figure 4a shows that *Stratified Winsorizing* on average results in a smaller treatment bias compared with the traditional approach to winsorizing for small and moderate treatment effects, ranging from Cohen's d $= [-0.5, 0.5]$ (Cohen, 1988). Figure 4b reproduces Figure 4a for larger treatment effects in the range of Cohen's d $= [-2, 2]$. While differences between the two approaches to winsorizing are not statistically significantly different (paired t-test), *Stratified Winsorizing* generates a smaller mean bias, smaller spread, and the bias does not increase or flip sign with

[9]The focus for this section is on winsorizing, however Appendix A reproduces simulations for trimming, with qualitatively similar results. The Appendix also reproduces the simulations for other distributions aside from a normal distribution for both winsorizing and trimming, with unchanged results.

the treatment effect (K-S test, $p < 0.001$). In cases of a positive treatment effect, the traditional approach to winsorizing can underestimate the treatment effect. When the treatment effect is negative, the traditional approach on average underestimates the true negative treatment effect by generating a positive bias on the treatment effect estimate.



(a) Cohen's d = $[-0.5, 0.5]$

(b) Cohen's d = $[-2, 2]$

(c) Varying Mean and S.D.

(d) Sample Size

Figure 4. Varying Normalized Treatment Effect

Figure 4c shows the effects of 10,000 random draws of the mean and standard deviation of the treatment and control distributions, with a range $(0, 4)$. *Stratified Winsorizing* outperforms the traditional approach to winsorizing, with a statistically insignificant smaller mean bias (paired t-test, p=0.492), but a statistically significantly smaller spread (K-S test, p<0.001). Figure 4d fixes the treatment effect to Cohen's d = 0.5, but varies the share of the sample belonging to the treatment group from 5% to 95%. Compared with the traditional approach to winsorizing, the bias arising from *Stratified Winsorizing* is consistent across the range of sample allocations, and smaller in magnitude. This difference in bias is highly statistically significant (paired t-test and K-S test, $p < 0.001$).

3.1.1 What is Driving These Results?

To understand the reduced treatment effect bias from *Stratified Winsorizing* compared with the traditional approach of winsorizing the whole sample, emphasis is placed on the observations that are winsorized, and the share of winsorized observations that are from the treatment and control group. *Stratified Winsorizing* ensures that a proportional share of observations are winsorized from the control and treatment groups. The simulations ensure that outliers are uncorrelated with treatment status, and thus the likelihood of an observation being winsorized should be uncorrelated with treatment status too. As treatment and control groups are equally sized in the simulations, proportional winsorizing would result in 50% of the winsorized observations being from the treatment group.

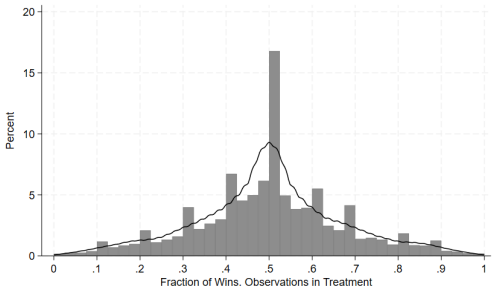
Figure 5a plots a histogram of the share of winsorized observations that are from the treatment group when using the traditional approach of winsorizing the whole sample. In some simulations, 100% of winsorized observations are from the treatment group, while in other simulations, 0% of winsorized observations are from the treatment group. In only 10.16% of the 10,000 simulations underlying Figure 4c does the traditional approach to winsorizing result in equal proportions of observations from the control and treatment group being winsorized.

Figures 5b and 5c plot the fraction of left- and right-tailed observations that are winsorized from the treatment group using the traditional approach to winsorizing and *Stratified Winsorizing*, as a function of the treatment effect size.[10] *Stratified Winsorizing* ensures that control and treatment groups are winsorized proportionately, irrespective of the size of the treatment effect. This results in 50% of winsorized observations being from the treatment group. The traditional approach to winsorizing, on the other hand, winsorizes control and treatment groups disproportionately. When treatment effects are negative, a larger share of left-tailed observations are winsorized from the treatment group, while a smaller share of right-tailed observations are winsorized from the treatment group, compared with the control group. When treatment effects are positive, the effect is reversed, and disproportionately more right-tailed observations are winsorized from the treatment group.

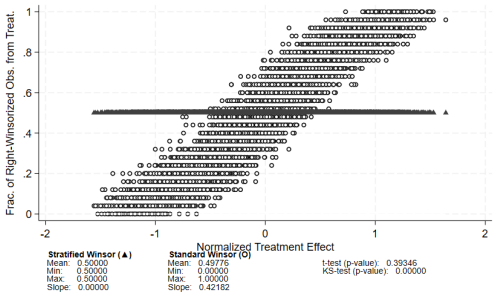
The intuition for these results can be traced back to Figure 2: the larger the treatment effect, the more right-tailed observations of the treatment group are win-

[10]The data is based on the 10,000 simulations underlying Figure 4b.

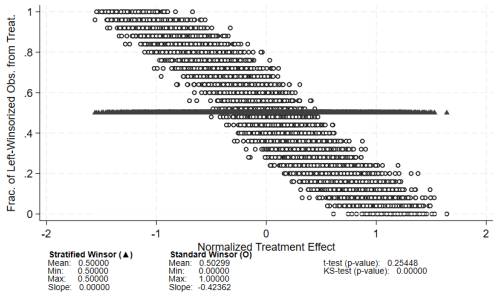
sorized when using the traditional approach to winsorizing, and the fewer left-tailed observations of the treatment group are winsorized. The line of best fit of the fraction of winsorized right-tailed observations from the treatment group has a slope of 0.42, implying that a 0.1 standard deviation increase in the treatment effect size results in the percentage of winsorized observations from the treatment group increasing by 4.2%.



(a) Fraction of Wins/ Observations in Treatment Group, Standard Wins. Approach



(b) Share of Left-Tailed Winsorized Observations from Treatment Group



(c) Share of Right-Tailed Winsorized Observations from Treatment Group

Figure 5. Monte Carlo Simulations: Winsorized Observations

Ensuring that a proportional share of observations are winsorized from the control and treatment groups also means that *Stratified Winsorizing* reduces the average distance of the winsorized variable from the nearest non-winsorized variable.[11] This can be explained by comparing Figures 3 and 4, which illustrate how the two approaches to winsorizing differ. *Stratified Winsorizing* ensures that only values

[11]For example, a distribution is winsorized at the 5th and 95th percentile. If an observation at the 99th percentile had an initial value of 10 (which would get winsorized), and an un-winsorized observation at the 95th percentile had a value of 5, the distance in absolute value would be $|10 - 5| = 5$.

greater than the 95th percentile of each subgroup's distribution are winsorized.[12] The traditional approach to winsorizing instead can result in values smaller than the 95th percentile of a subgroup's distribution getting winsorized, which increases the distance between the value of the winsorized and non-winsorized observations.[13] In the simulations underlying Figure 4c, *Stratified Winsorizing* reduced the average distance of a winsorized from a non-winsorized observation by 8.03%, compared with the traditional approach to winsorizing. This difference in distance between the two winsorizing techniques is highly statistically significantly ($p<0.001$, paired t-test and K–S test, see Appendix A.1.4).

3.2 Type II Errors and Statistical Power

The effects of both approaches to winsorizing can affect the likelihood of Type II errors, and thus a study's statistical power. To identify this, 1000 iterations are run, each consisting of 1000 simulations of an RCT with 500 observations. In each iteration, the sample size is 500 subjects, equally divided across treatment and control groups. Two-sided t-tests of independent observations are performed, with a significance level of $\alpha = 0.05$. The control group is characterized by a standard normal distribution, while the treatment group is a normal distribution with a standard deviation of 1, but a non-zero mean. Additionally, the outcome variable includes a standard normal error term. The resulting distributions are winsorized at the 90% level (top and bottom 5%), using both winsorizing techniques.

For each iteration, statistical power is calculated as the percentage of simulations in which the treatment effect is statistically significant. This is performed separately for the whole sample, and the winsorized sample using the traditional approach, and *Stratified Winsorizing*. Figure 6 reports the percentage improvements in the study's statistical power as a result of the two approaches to winsorizing, compared with no winsorizing.

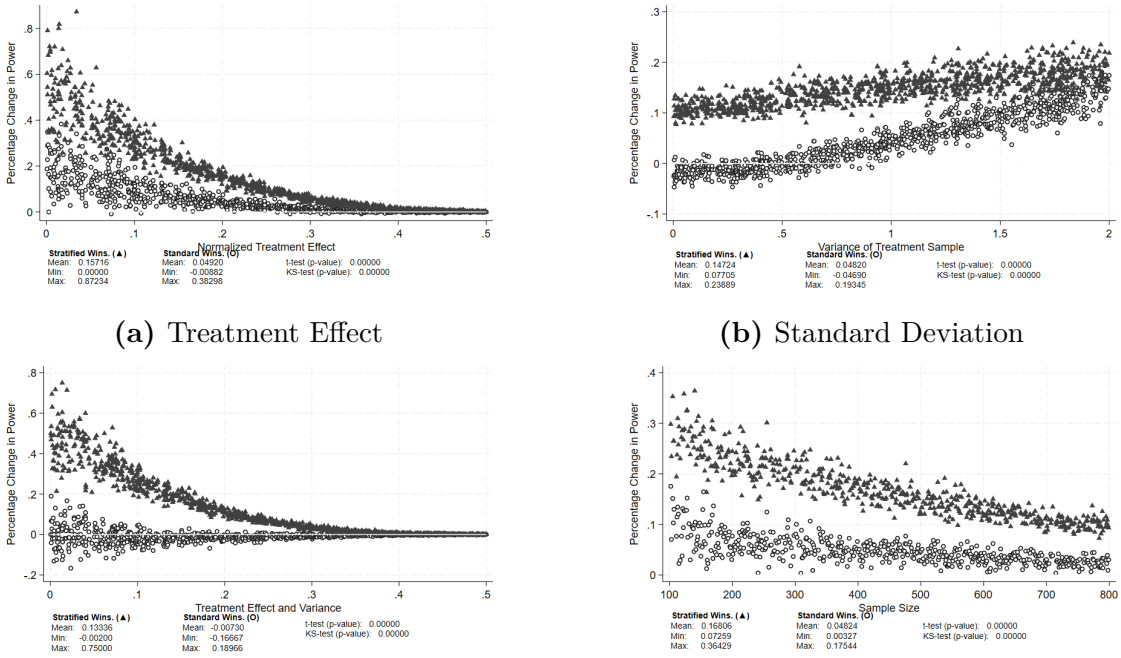
For Figure 6a, the treatment effect is a uniformly drawn value between $d = [0, 0.5]$. In Figure 6b, the treatment effect is Cohen's d=0.2, while the variance of

[12] The focus here is on the right tail, however the intuition is identical for the left tail (5th percentile).

[13] For example, if Figure 1 simulates winsorizing the sample at the 5th and 95th percentile, then Figure 2 showcases that the traditional approach to winsorizing would winsorize the 10th percentile and below of the Control group, and the 90th percentile and above of the treatment group. The assumption here is that control and treatment have an equal sample size.

the treatment's normal distribution varies uniformly between 0 and 2. In Figure 6c, the mean equals the variance of the treatment group's distribution, and is a value between $(0, 0.5]$. Figure 6d keeps the treatment group's distribution fixed ($\sim N(2, 1)$), but varies the sample size of the distribution from 100 to 800 (with the sample being evenly split between treatment and control group).

What is consistent across Figure 6 is that *Stratified Winsorizing* outperforms the traditional winsorizing technique, in terms of statistical power and hence the likelihood of Type II errors, particularly in simulations with a small treatment effect or small sample size, which typically have lower levels of statistical power.



(a) Treatment Effect

(b) Standard Deviation

(c) Mean and Standard Deviation

(d) Sample Size

Figure 6. Effects of Winsorizing on Statistical Power

3.3 Type I Errors

The null hypothesis of the simulated RCT regression is that $\beta_1 = 0$, hence that treatment and control groups are from the same underlying distribution. With a significance level $\alpha = 0.05$, the expectation is that Type I errors – where the null hypothesis of no treatment effect is incorrectly rejected – occur in 5% of the

cases. A concern with *Stratified Winsorizing* is that Type I errors can emerge with a greater likelihood if the researcher assumes that the sample distribution consists of subgroups, while in fact it does not. In that case, winsorizing per subgroup can lead to distortions, and increase the likelihood of Type I errors.

Table 1 reports the likelihood with which Type I errors occur. Results are based on 1000 iterations, each consisting of 1000 simulations of the RCT with 500 observations. The control and treatment groups are drawn from the same distribution, and hence the true treatment effect is zero. Two-sided t-tests of independent observations are performed to estimate treatment effects, with a significance level of $\alpha = 0.05$. Therefore, Type I errors are expected in 5% of the cases. Simulations are conducted for normal, log-normal, skew-normal, and gamma distributions.

As Table 1, Panel A illustrates, *Stratified Winsorizing* increases the probability of Type I errors in instances where the sample distribution is not composed of subgroups. While the frequency of Type I errors is not statistically significantly different when outliers are not winsorized compared to when the whole sample is winsorized, *Stratified Winsorizing* results in statistically significantly more cases of Type I errors.

Panel B of Table 1 uncovers an interesting dynamic: while the likelihood of Type I errors is higher when using the *Stratified Winsorizing* technique, the likelihood of a Type I error when there is no winsorizing also being a Type I error when winsorizing is greater using the *Stratified Winsorizing* than the traditional approach of winsorizing the entire sample. The observation that not all of the same Type I errors are documented when winsorizing vs. not is in line with Bollinger and Chandra (2005), who argue that the remaining sample after winsorizing differs from the sample without winsorizing. This can affect not only the treatment effect estimates (and hence Type II errors) but also the likelihood of Type I errors.

4 Applications to Angelucci et al. (2023) and Jack et al. (2023)

The Monte Carlo simulations demonstrate that both approaches to winsorizing can affect a study's estimated treatment effect, and the likelihood of Type I and II errors. In this Section, I illustrate how the two approaches to winsorizing/trimming

Table 1: Winsorizing and Type I Errors

	Normal Distr.	Log-Normal Distr.	Skew-Normal Distr.	Gamma Distr.
A. Frequency of Type I errors				
No Winsor	0.050	0.048	0.050	0.050
Traditional Wins.	0.050	0.050	0.050	0.050
Stratified Wins.	0.075	0.108	0.069	0.081
p-value No vs. Trad.	0.28	0.00	0.68	0.90
p-value No vs. Strat.	0.00	0.00	0.00	0.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00
B. Percentage of *No Winsor* Type I errors included				
Traditional Wins.	85.24	61.90	88.14	80.98
Stratified Wins.	99.18	92.78	99.25	98.37
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00

can affect the statistical significance of treatment effect estimates, using Angelucci et al. (2023) and Jack et al. (2023) as examples. Appendix D performs similar analysis on Schilbach (2019) and Augsburg et al. (2015). These studies were chosen due to their different types of data (administrative vs. self-reported), monetary and non-monetary outcomes, different empirical strategies (RCT vs. Difference-in-Differences), uses of trimming and winsorizing, and the availability of their data and code. The regression tables first replicate the findings of the respective paper – using the traditional approach to winsorizing/trimming – in Panel A, followed by regression estimates using the *Stratified Winsorizing/Trimming* technique in Panels B and C. Panel B stratifies by treatment group, while Panel C also stratifies by the data collection round. Panels D-F illustrate the percentage of observations winsorized/trimmed for each of the subgroups and survey rounds using the different winsorizing/trimming techniques.

4.1 Angelucci et al. (2023, JDE)

Angelucci et al. (2023) conducted an RCT among women in the Democratic Republic of Congo, randomizing access to a multifaceted program including financial support, training, and social support. The study measured the intervention's impact on various outcomes, immediately after the program ended (endline), and one

year later (follow-up). Data was winsorized at the 5th and 95th percentiles.[14] Table 2.(i) reports the OLS-estimated treatment effects for Total Monthly Earnings, Earnings Net of Costs, and Total Business Costs.[15] Panel A replicates the findings of Table 4 in Angelucci et al. (2023) by using the traditional winsorization technique, while Panels B and C present OLS regression results for the *Stratified Winsorizing per Treatment* and *Stratified Winsorizing per Treatment*TimePeriod* approaches, respectively.

Table 2 reports treatment effects of the intervention for both the endline survey (after the end of the intervention), and the follow-up (one year later). Compared with Panel A, *Stratified Winsorizing by Treatment*, and *by Treatment*TimePeriod* (Panels B and C, respectively) result in larger treatment effect estimates, with greater statistical significance. This suggests that the traditional approach to winsorizing has a downward bias on the treatment effect estimates.

Table 2.(ii) illustrates that this downward bias is driven by an over-winsorizing of right-tailed observations from the treatment group, as it reports the percentage of observations winsorized in the treatment and control groups of Angelucci et al. (2023), as well as the percentage of observations winsorized at endline and the post-endline follow-up. Panel D demonstrates that the traditional approach to winsorizing differentially winsorizes control and treatment observations, with a greater percentage of treated observations being winsorized than observations in the control group. The discrepancy between the percentage of observations winsorized in the control and treatment group is reduced as a result of *Stratified Winsorizing by Treatment*, as shown in Panel E.

However, Table 2.(ii) also illustrates that the traditional winsorizing approach and *Stratified Winsorizing by Treatment* technique differentially winsorize observations from different survey rounds. Both techniques winsorize endline observations more than 1-year follow-up observations – although it is unlikely that the measurement error was systematically higher during the endline survey. This is addressed by Panels C and F, which winsorize the data stratified by *Treatment*TimePeriod*, to further ensure that not only are observations from different treatment groups

[14]Nevertheless, only right-tailed observations are winsorized. This is because for all three outcome variables, over 50% of observations equaled 0, the lower bound. Hence no winsorizing took place at the left tail.

[15]These outcome variables were chosen, as they were the only ones that were winsorized in the replication package.

winsorized proportionately, but also across survey rounds.

Table 2: Angelucci et al. (2023), Table 4

Table 2.(i) OLS Treatment Effect Estimates

	Total Monthly Earnings		Earnings Net of Costs		Total Business Costs	
	Endline	Follow-up	Endline	Follow-up	Endline	Follow-up
	(1)	(2)	(3)	(4)	(5)	(6)
A. Traditional Winsorizing						
Treatment	0.202*	0.467***	0.0714	0.191**	0.180**	0.321***
	(0.106)	(0.120)	(0.0704)	(0.0773)	(0.0731)	(0.0859)
B. Stratified Winsorizing by Treatment						
Treatment	0.365***	0.585***	0.146**	0.263***	0.429***	0.577***
	(0.114)	(0.118)	(0.0727)	(0.0768)	(0.0771)	(0.103)
C. Stratified Winsorizing by Treatment*TimePeriod						
Treatment	0.309***	0.681***	0.166**	0.249***	0.301***	0.635***
	(0.112)	(0.126)	(0.0699)	(0.0776)	(0.0672)	(0.107)

Table 2.(ii) % of Treat. and Control Obs. Winsorized

	Total Monthly Earnings	Earnings Net of Costs	Total Business Costs
	(1)	(2)	(3)
D. Traditional Winsorizing			
% of Control Obs. Winsorized	2.95	5.45	3.25
% of Treatment Obs. Winsorized	5.72	9.82	5.82
% of Endline Obs. Winsorized	4.75	7.94	4.75
% of Follow-up Obs. Winsorized	3.97	7.40	4.36
E. Stratified Winsorizing by Treatment			
% of Control Obs. Winsorized	4.65	7.15	4.45
% of Treatment Obs. Winsorized	4.62	8.66	4.43
% of Endline Obs. Winsorized	4.95	8.04	4.60
% of Follow-up Obs. Winsorized	4.31	7.79	4.26
F. Stratified Winsorizing by Treatment*TimePeriod			
% of Control Obs. Winsorized	3.90	7.15	4.30
% of Treatment Obs. Winsorized	4.57	8.71	4.52
% of Endline Obs. Winsorized	4.41	8.19	4.56
% of Follow-up Obs. Winsorized	4.07	7.70	4.26

Notes: Standard errors are in parentheses, and clustered at the level of the treatment group. Stratified Winsorizing by Treatment winsorizes the sample separately for treatment and control, while Traditional Winsorizing winsorizes the entire sample. Stratified Winsorizing by Treatment*TimePeriod winsorizes the sample separately for treatment and control observations at endline and follow-up separately. Results are reported without corrections for multiple hypothesis testing. Variables are winsorized at the 5th and 95th percentiles. Consumption refers to the previous week. Business costs include the discounted use value of large purchases. * p<0.1, ** p<0.05, *** p<0.01

Compared with Panel A, treatment effects reported in Panel C are larger in magnitude, and statistically more significant. This is driven by the winsorized observations being evenly distributed across treatments, and survey rounds, as shown in Table 2.(ii). Panels C and F highlight the importance of not only stratifying winsorizing by treatment, but also by the survey round - particularly for empirical strategies where the outcome variable is measuring at different time periods. This will be discussed more in the application to Jack et al. (2023) and the practical implications in Section 5.

4.2 Jack et al. (2023)

Jack et al. (2023) conducted an RCT among Kenyan farmers and offered four different loan offers to purchase a water harvesting tank, with varying degrees of asset collateralization. To measure the intervention's impact on milk sales based on administrative data, the researchers use a ITT difference-in-differences approach, and trim the data at the 1, 5, and 10% level (only the right tail) to account for outliers.

Table 3.(i), Panel A reproduces Table 6 of Jack et al. (2023) by reporting treatment effects using the traditional approach to trimming, while Panel B reports treatment effects using the *Stratified Trimming by Treatment* technique. Panel B reports larger and more statistically significant treatment effects than Panel A, suggesting that the traditional approach to trimming can have a downward bias on the treatment effect estimate.

As Table 3.(ii) demonstrates, the traditional approach to trimming results in differential trimming of observations in treatment and control groups. In line with the intervention having a positive treatment effect and the authors only trimming the right-hand tail, Panel D illustrates that a disproportionately larger share of treatment group observations get trimmed using the traditional approach to trimming. Panel E shows that this is overcome using the *Stratified Trimming by Treatment* technique.

Table 3.(ii) also illustrates that both the traditional approach of trimming the whole sample and *Stratified Trimming by Treatment* techniques differentially trim between baseline and endline observations in Jack et al. (2023). Both techniques trim endline observations more than baseline observations – despite it being unlikely that outliers were systematically more common at endline.

Panel C in Table 3.(i) reports the OLS treatment effect estimates when using *Stratified Trimming by Treatment*TimePeriod*, ensuring that a proportional share of baseline and endline observations are trimmed, both for treatment and control groups. Compared with Panel A, the *Treat*Post* OLS estimate of the treatment effects increases by 21%, 10%, and 27% (for 1%, 5%, 10% trimming, respectively). These changes in magnitude are large, statistically significant, and driven by the trimmed observations being evenly distributed across treatments, and time periods, as illustrated in Panel F.

5 Practical Guidelines

The Monte Carlo simulations and applications to Angelucci et al. (2023) and Jack et al. (2023) have illustrated that the decision of how to winsorize/trim observations to reduce the role of outliers is less innocuous than it initially seems and can have large effects on the treatment effect estimates. The Monte Carlo simulations further showed that the chosen winsorizing/trimming technique can affect the likelihood of Type I and II errors. This Section therefore discusses practical guidelines when considering whether and how to winsorize/trim outliers.

5.1 When to Use Which Technique

The underlying empirical strategy and data generating process should inform the decision of whether and how to winsorize/trim. Regarding the first decision of whether to winsorize/trim, if outliers persist across correlated outcome variables, it is unlikely these outliers are due to repeated measurement errors, and more likely represent a large treatment effect for a few observations. When treatment effects are driven by these sorts of outliers that are not due to measurement errors – for example the large effects of microcredit among the upper tails across seven studies reported by Meager (2022) – winsorizing these outliers will bias the true treatment effect. In these cases, complementing average treatment effects with quantile regressions can highlight the overall effect of the intervention as well as its heterogeneity.

The second decision is how to winsorize/trim. In cases where a value beyond/below a certain value can easily be identified as outliers (e.g., the upper bound of the WTA measure of Allcott et al. (2020)), authors should consider those observations outliers and winsorize/trim them accordingly. However, the majority of academic papers set arbitrary percentile thresholds (e.g., 99th or 95th percentile). In these cases, the decision on whether to winsorize/trim the whole sample or separately per subgroup should depend on the empirical strategy deployed.

Irrespective of the empirical strategy, **panel data collected during different time periods/survey rounds should be treated as separate subgroups, and hence winsorized/trimmed separately.** As the examples of Angelucci et al. (2023) and Jack et al. (2023) illustrate, observations of certain time periods are more likely to be winsorized/trimmed when winsorizing/trimming is not done

separately per time period, despite no clear rationale existing for why outliers are more common in certain time periods. As such, it is important to winsorize/trim observations from each time period / survey wave separately.

Both winsorizing/trimming techniques have their advantages and disadvantages, as the Monte Carlo simulations illustrated. While *Stratified Winsorizing/Trimming* can improve a study's statistical power and reduce the bias of treatment effect estimates, it can increase the likelihood of Type I errors compared with the traditional approach of winsorizing/trimming the whole sample when the underlying distribution is drawn from the same sample. **With Randomized Controlled Trials, there is no clear winner. Instead, reporting both techniques can provide a more robust estimation of the treatment effect, while minimizing the effects of Type I and II errors.** This is because the underlying null hypothesis of RCTs is that treatment and control groups are drawn from the same distribution. Reporting treatment effects using both winsorizing/trimming techniques can strengthen the robustness of the treatment effect by illustrating that outliers are not driving the treatment effects, in line with the insights of Young (2019) and Broderick et al. (2023).

When treatment effects differ substantially as a result of the winsorizing/trimming technique used, it is important to understand why. For this, an understanding of the underlying data generating process is crucial: if differential winsorizing/trimming of subgroups is observed when winsorizing/trimming the whole sample (like in Panel D of the applications to Angelucci et al. (2023) and Jack et al. (2023)), a justification is needed. For example, if experimenter demand effects are stronger among a certain subgroup, it can be justified to disproportionately winsorize/trim outliers from that subgroup. However, without a clear rationale why subgroups are disproportionately winsorized/trimmed, *Stratified Winsorizing/Trimming* is likely to report treatment effect estimates closer to the true treatment effect by ensuring that equal proportions of observations are winsorized/trimmed across the subgroups.

In cases where *Stratified Winsorizing/Trimming* results in statistically significant estimates but the traditional approach to winsorizing/trimming does not, authors need to be careful that the statistically significant treatment effect estimates as a result of *Stratified Winsorizing/Trimming* are not due to an increased likelihood in Type I errors. The Monte Carlo simulations illustrated that Type I errors are more likely as a result of the *Stratified Winsorizing/Trimming* technique. In such

cases, the recommendation is for the authors to report treatment effect estimates following the traditional approach to winsorizing/trimming, in order to minimize the risks associated with Type I errors. Only if authors can justify why treatment effect estimates of *Stratified Winsorizing/Trimming* are more likely to be reliable (e.g., differential winsorizing/trimming of subgroups using the traditional approach to winsorizing/trimming although there is no clear reason why), should they be reported as main results.

This recommendation differs for other kinds of empirical strategies. Unlike RCTs, subgroups in DiD and RDD specifications are not randomized from the same underlying sample. Instead, they are drawn from different samples. As such, differences between these subgroups are likely to be more pronounced, increasing the likelihood that winsorizing/trimming the whole sample will differentially winsorize/trim observations across subgroups. For example, with Difference-in-Difference designs, the distributions of "treatment" and "control" can be very different so long as the parallel trends assumption is satisfied. Similarly, with Regression Discontinuity Designs, "treatment" and "control" observations are drawn from different sides of a threshold cutoff. As such, the major drawback of *Stratified Winsorizing/Trimming* – namely the increased likelihood of Type I errors – primarily applies to RCTs. Therefore, **the recommendation is to use the *Stratified Winsorizing/Trimming* for Difference-in-Difference and Regression Discontinuity Designs**, to minimize the likelihood of differential winsorizing/trimming of treatments and hence the resulting Type II errors.

5.2 Pre-Analysis Plans

While the data generating process should inform the decision of how to deal with outliers, the rise of Pre-Analysis Plans means that authors have to announce their strategy for dealing with outliers before understanding the underlying data generating process. Of all the Stage I accepted Pre-Analysis Plans at the Journal of Development Economics that indicated their intention to winsorize/trim their data, all bar one winsorize/trim their data at either the 95th or 99th percentile.[16] For future Pre-Analysis Plans of RCTs, a recommendation is to **pre-specify that**

[16]Only Angelucci and Bennett (2024) do not winsorize/trim at the 95th or 99th percentile, and instead winsorize observations outside 1.5 times the inter-quartile range, following the suggestion of Beyer (1981).

both approaches to winsorizing/trimming will be used as a pre-specified percentile cut-off, in order to provide further robustness that treatment effect estimates are not driven by outliers.

For papers without Pre-Analysis Plans, a documentation of how outliers are handled, including which winsorizing/trimming threshold and technique are chosen, in the paper's appendix will increase the transparency surrounding data cleaning and analysis. In addition to this documentation, **reporting the proportion of winsorized/trimmed observations per subgroup – like Tables 2.(ii) and 3.(ii) – illustrates whether sub-groups are disproportionately affected.** If both winsorizing/trimming approaches are used, reporting how the proportion of winsorized/trimmed observations per subgroup differs by winsorizing/trimming approach can explain differences in observed treatment effects.

5.3 How to Define Subgroups

When stratifying winsorizing/trimming by subgroups, authors need to decide how to define sub-groups. For RCTs, different treatment arms should be considered as subgroups, along with different survey waves, as shown in the application to Angelucci and Bennett (2024).[17] For Difference-in-Differences empirical strategies, sub-groups should be stratified on survey waves, and treatment groups, as illustrated by Jack et al. (2023). The same holds for regression discontinuity designs.

A concern arises when too many subgroups are defined: akin to stratified randomization, if authors define too many stratas/subgroups, each subgroup will be so small that no outliers get winsorized/trimmed. Furthermore, creating subgroups when there in fact are no subgroups can increase the likelihood of Type I errors, as the Monte Carlo simulations illustrated. Finally, defining too many subgroups can complicate the interpretability of treatment effects across regression tables: for example, if an author of an RCT defines subgroups differently for the main regression (comparing treatment and control) and gender heterogeneity regressions — by defining subgroups as *Treatment*Gender* in the second regression — treatment effect estimates between the two regressions are harder to compare as the observations that are winsorized/trimmed differ between the two regressions. Therefore, the rec-

[17] At baseline, treatment arms should not be winsorized/trimmed separately, because randomization should ensure they are from the same underlying distribution.

ommendation is to **define subgroups by time periods (in the case of panel data), and "treatment" groups.**

5.4 Statistical Software

Below, the code for the traditional and stratified approach to winsorizing can be found for Stata and R. Online Appendix C shows the code for the traditional and stratified approach to trimming.

5.4.1 Stata

Traditional approach to winsorizing: winsor2 *OutcomeVar*, cuts(5 95)
Stratified Winsorizing: winsor2 *OutcomeVar*, cuts(5 95) by(*StratifiedVariable*)

5.4.2 R

I developed a new R package, called WinsorByGroupR, which can be found on GitHub. Once the package is installed, the functions are as follows:
Traditional approach to winsorizing: winsor(data, value_col = "OutcomeVar", bounds = c(5, 95))
Stratified winsorizing: winsorize_by_group(data, group_col = "StratifiedVariable", value_col = "OutcomeVar", bounds = c(5, 95))

6 Conclusion

Winsorizing and trimming are frequently used to reduce the role of outliers in dependent variables, by defining a percentile beyond which observations are considered outliers and hence winsorized/trimmed. However, this paper illustrates that winsorizing and trimming is less innocuous than it seems and can bias a study's treatment effect estimates. These findings are in line with findings by Broderick et al. (2023) and Young (2019), who show that a few observations can have large effects on treatment effect estimates. This paper further shows how the winsorizing/trimming technique used can affect the likelihood of Type I and Type II errors.

While most papers winsorize/trim the entire sample, recent studies — including Benson et al. (2023), Muralidharan et al. (2023), and Bedoya et al. (2023)

— have winsorized/trimmed separately per subgroup, a technique called *Stratified Winsorizing/Trimming*. Monte Carlo simulations of an RCT illustrate that *Stratified Winsorizing/Trimming* on average result in a smaller bias of the treatment effect estimate, compared with the traditional approach of winsorizing/trimming the whole sample. Furthermore, *Stratified Winsorizing/Trimming* improved the study's statistical power, at the cost of increasing the likelihood fo Type I errors.

Applications to Angelucci et al. (2023) and Jack et al. (2023) illustrate that the decision of *how* to winsorize/trim has empirical implications, as treatment effects and their statistical significance change. As such, authors should carefully consider how to winsorize/trim outliers, informed by the underlying data generating process.

The focus of the simulations and empirical applications has been on RCTs – given those are the most common empirical setting in which outliers are winsorized/trimmed – however, the insights and implications also translate to other empirical approaches such as Difference-in-Difference or Regression Discontinuity Designs.

Table 3: Jack et al. (2023), Table 6

Table 3.(i) OLS Treatment Effect Estimates

	(1) Milk Sales 1% trim	(2) Milk Sales 5% trim	(3) Milk Sales 10% trim
A. Traditional Trimming			
Treat*Post	12.580*	12.749**	9.790**
	[6.419]	[5.106]	[4.389]
Treatment	-3.568	-5.960	-6.161
	[5.804]	[4.691]	[3.914]
B. Trimming by Treatment			
Treat*Post	13.355**	14.374***	11.320***
	[6.404]	[5.053]	[4.339]
Treatment	-1.832	-5.172	-4.374
	[5.640]	[4.653]	[3.847]
C. Trimming by Treatment*TimePeriod			
Treat*Post	15.219**	14.061***	12.398***
	[6.415]	[5.091]	[4.271]
Treatment	-3.360	-4.935	-4.890
	[5.258]	[3.942]	[3.070]

Table 3.(ii) % of Treat. and Control Obs. Trimmed

	(1) 1% trim	(2) 5% trim	(3) 10% trim
D. Traditional Trimming			
% of Control Obs. Trimmed	0.79	4.63	9.26
% of Treatment Obs. Trimmed	1.06	5.11	10.21
% of Baseline Obs. Trimmed	0.50	2.38	5.14
% of Endline Obs. Trimmed	1.10	5.53	10.98
E. Stratified Trimming by Treatment			
% of Control Obs. Trimmed	1.00	4.99	9.96
% of Treatment Obs. Trimmed	1.00	5.00	9.99
% of Baseline Obs. Trimmed	0.52	2.45	5.17
% of Endline Obs. Trimmed	1.10	5.53	10.96
F. Stratified Trimming by Treatment*TimePeriod			
% of Control Obs. Trimmed	0.99	4.98	9.99
% of Treatment Obs. Trimmed	0.99	4.99	9.99
% of Baseline Obs. Trimmed	0.99	4.98	9.97
% of Endline Obs. Trimmed	1.00	4.99	9.99

Notes: The Post dummy refers to all months from June 2010 (the median loan offer date) onwards. Milk sales are reported in liters. A 1% trim means the top percentile of observations have been trimmed; similarly for the 5% and 10% trims. Standard errors clustered at household level are reported in brackets. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author(s) used ChatGPT Deep Research in order to undergo a simulated review process prior to submission. Furthermore, ChatGPT was used for the coding of the Monte Carlo simulations. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–76.

Angelucci, M. and Bennett, D. (2024). Pharmacotherapy and weight loss in india: A pre-analysis plan. Pre-Analysis Plan, accessed: 2025-02-21.

Angelucci, M., Heath, R., and Noble, E. (2023). Multifaceted programs targeting women in fragile settings: Evidence from the Democratic Republic of Congo. *Journal of Development Economics*, 164:103146.

Angrist, J. D. and Krueger, A. B. (2000). *Empirical Strategies in Labor Economics*, volume 3A, chapter 23, pages 1277–1366. Elsevier Science, Amsterdam.

Augsburg, B., De Haas, R., Harmgart, H., and Meghir, C. (2015). The Impacts of Microcredit: Evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1):183–203.

Bedoya, G., Belyakova, Y., Coville, A., Escande, T., Isaqzadeh, M., and Ndiaye, A. (2023). The Enduring Impacts of a Big Push during Multiple Crises: Experimental Evidence from Afghanistan. Technical report, World Bank Group.

Benson, A., Board, S., and Meyer-ter Vehn, M. (2023). Discrimination in Hiring: Evidence from Retail Sales. *The Review of Economic Studies*, 91(4):1956–1987.

Beyer, H. (1981). Tukey, john w.: Exploratory data analysis. addison-wesley publishing company reading, mass. — menlo park, cal., london, amsterdam, don mills, ontario, sydney 1977, xvi, 688 s. *Biometrical Journal*, 23(4):413–414.

Bollinger, C. and Chandra, A. (2005). Iatrogenic Specification Error: A Cautionary Tale of Cleaning Data. *Journal of Labor Economics*, 23(2):235–258.

Broderick, T., Giordano, R., and Meager, R. (2023). An automatic finite-sample robustness metric: When can dropping a little data make a big difference?

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2 edition.

de Mel, S., McKenzie, D., and Woodruff, C. (2019). Labor drops: Experimental evidence on the return to additional labor in microenterprises. *American Economic Journal: Applied Economics*, 11(1):202–35.

Fafchamps, M., McKenzie, D., Quinn, S., and Woodruff, C. (2012). Using pda consistency checks to increase the precision of profits and sales measurement in panels. *Journal of Development Economics*, 98(1):51–57. Symposium on Measurement and Survey Design.

Goldberger, A. S. (1981). Linear regression after selection. *Journal of Econometrics*, 15(3):357–366.

Gollin, D. and Udry, C. (2021). Heterogeneity, Measurement Error, and Misallocation: Evidence from African Agriculture. *Journal of Political Economy*, 129(1):1–80.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1):153–161.

Heckman, J. (1990). Varieties of Selection Bias. *American Economic Review*, 80(2):313–318.

Jack, W., Kremer, M., de Laat, J., and Suri, T. (2023). Credit Access, Selection, and Incentives in a Market for Asset-Collateralized Loans: Evidence From Kenya. *The Review of Economic Studies*, 90(6):3153–3185.

Meager, R. (2022). Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 112(6):1818–47.

Muralidharan, K., Niehaus, P., and Sukhtankar, S. (2023). General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence From India. *Econometrica*, 91(4):1261–1295.

Schilbach, F. (2019). Alcohol and Self-Control: A Field Experiment in India. *American Economic Review*, 109(4):1290–1322.

World Bank (2023). Variable construction. https://dimewiki.worldbank.org/Variable_Construction. Accessed: 2024-02-23.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2):557–598.

Online Appendix to:

Winsorizing and Trimming with Subgroups

by Till Wicker

A Simulations

A.1 Biased Treatment Effect

A.1.1 Description

10,000 simulations are run of a hypothetical RCT with 500 subjects, that are divided equally across treatment and control (except for Figure 4d). The outcome variable is normally distributed and has a normally distributed error term ($\sim N(0,1)$). In Figures 4a and 4b, the normal distributions of the outcome variable of the treatment and control groups are characterized by a mean that is uniformly, randomly drawn from [0,0.5] ([0,2] for Figure 4b), with a standard deviation of 1.

The resulting distributions are winsorized at the 90% level (top and bottom 5%), using the traditional winsorizing approach, as well as the *Stratified Winsorizing per Treatment* approach. Outcome variable y (unwinsorized, traditional winsorizing, *Stratified Winsorizing per Treatment*) is then regressed on $Treatment$, with HAC robust standard errors ($y_i = \beta_1 T_i + \varepsilon_i$). Hence each simulation generates a treatment effect without winsorizing, and the two approaches to winsorizing. The resulting bias is measured as the difference in treatment effects (between the non-winsorized sample, and the winsorized sample, done separately for the two approaches to winsorizing), normalized by the standard deviation of the un-winsorized control group.

Figure 4c varies the mean and standard deviation of the treatment and control groups, with each taking a randomly and independently chosen value between 0 and 4. In Figure 4d, the control group is characterized by a standard normal distribution, while the treatment group is a normal distribution with mean 0.5 and standard deviation 1. Among the sample of 500 subjects, a random number between $[20, 480]$ is assigned to the treatment group.[18]

For Figure 7a, the standard deviations of the outcome variable of the treatment and control group are randomly and uniformly chosen values between 0 and 4. The mean of the treatment group's distribution is 3, while it is equal to 1 in the Control group (and

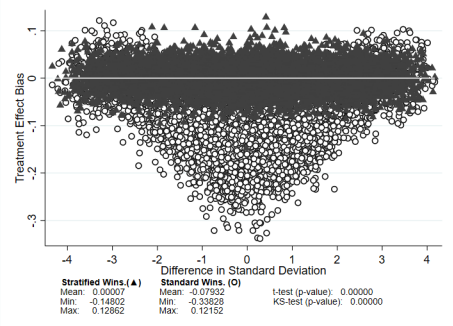
[18] Each treatment group needed at least 20 subjects such that trimming at the 90% level would winsorize at least one observation on each tail.

hence the average treatment effect =2). Because the standard deviation of the control group varies and can be very close to zero, the biases are not normalized, to avoid very large values.

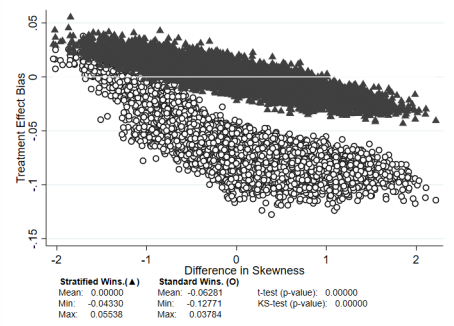
For Figure 7b, python's *skewnorm* function is used to simulate non-normal and non-symmetric distributions with skewness values ranging from -4 (left-tailed) to 4 (right-tailed), while keeping the mean and standard deviation constant. The same simulations are done with trimming (Figure 9), where the top and bottom 5% of the distribution are trimmed, rather than winsorized.

For non-normal distributions (Figures 8 and 10), 10,000 simulations were run, where the mean and standard deviation were a randomly drawn value between (0,4). In the case of the Poisson distribution, $\lambda \in (0, 4)$.

A.1.2 Winsorizing, Normal Distribution



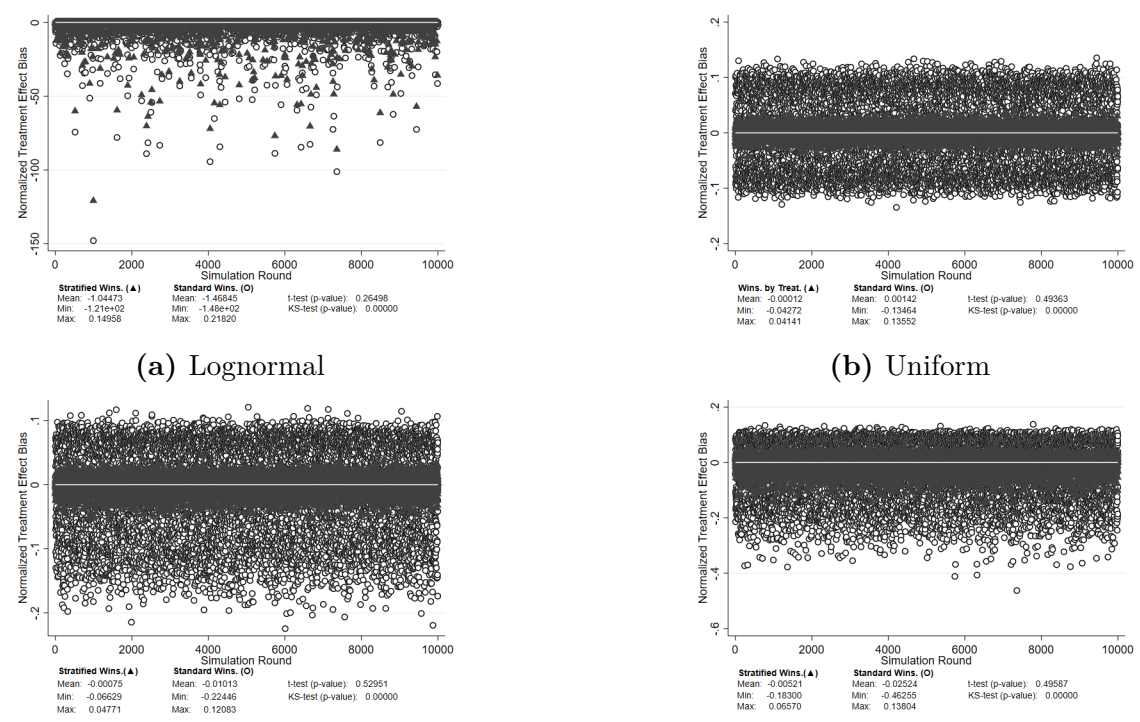
(a) Standard Deviation



(b) Skewness

Figure 7. Winsorizing: Normal Distribution

A.1.3 Winsorizing, Non-Normal Distribution



(a) Lognormal **(b)** Uniform

(c) Poisson **(d)** Gamma

Figure 8. Winsorizing: Non-normal Distributions

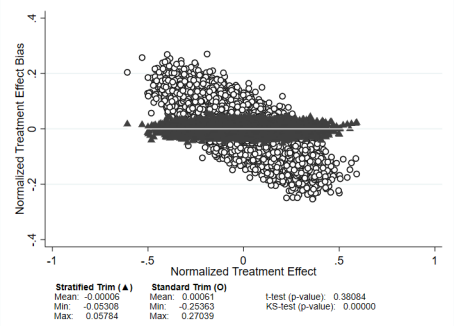
A.1.4 Winsorizing, Share of Winsorized Observations

Table 4: Winsorized Variables - Both Tails

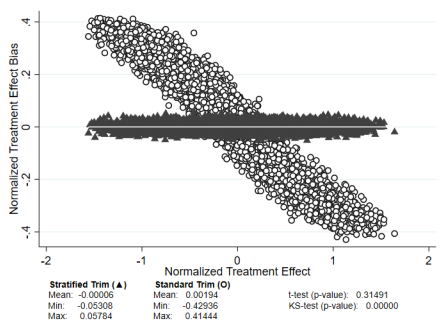
	Traditional Approach To Trim.	Stratified Trimming	Paired t-test p-value	KS-Test p-value
A. Average Distance from Non-Winsorized Value				
Entire Sample	0.772	0.710	0.000	0.000
	(0.001)	(0.001)		
Control	0.730	0.710	0.000	0.000
	(0.002)	(0.002)		
Treatment	0.731	0.711	0.000	0.000
	(0.002)	(0.002)		
B. Share of Winsorized Observations from Treatment				
Treatment	0.500	0.500	0.988	0.000
	(0.002)	(0.000)		

Notes: Standard errors are reported in brackets. Data is based on the 10,000 simulations underlying Figure 4c.

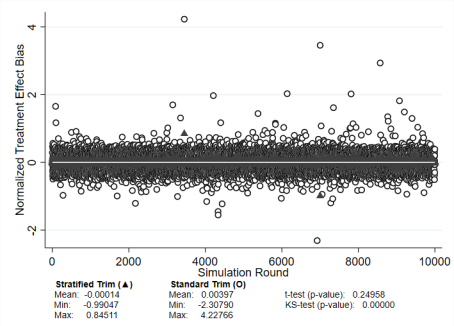
A.1.5 Trimming, Normal Distribution



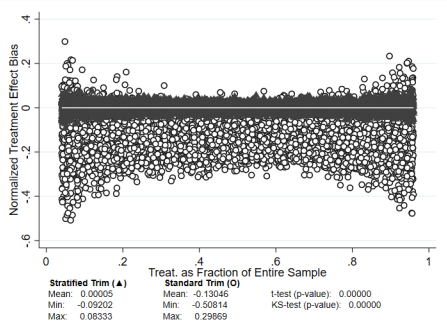
(a) Cohen's d = $[-0.5, 0.5]$



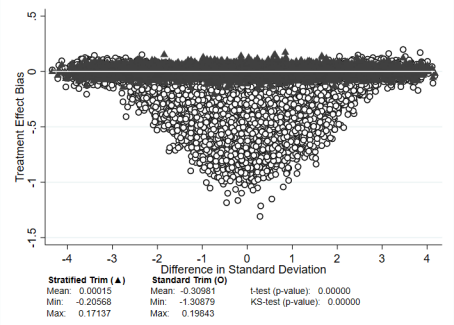
(b) Cohen's d = $[-2, 2]$



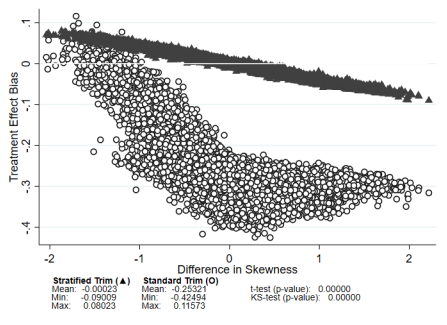
(c) Varying Mean and S.D.



(d) Varying Treatment Share



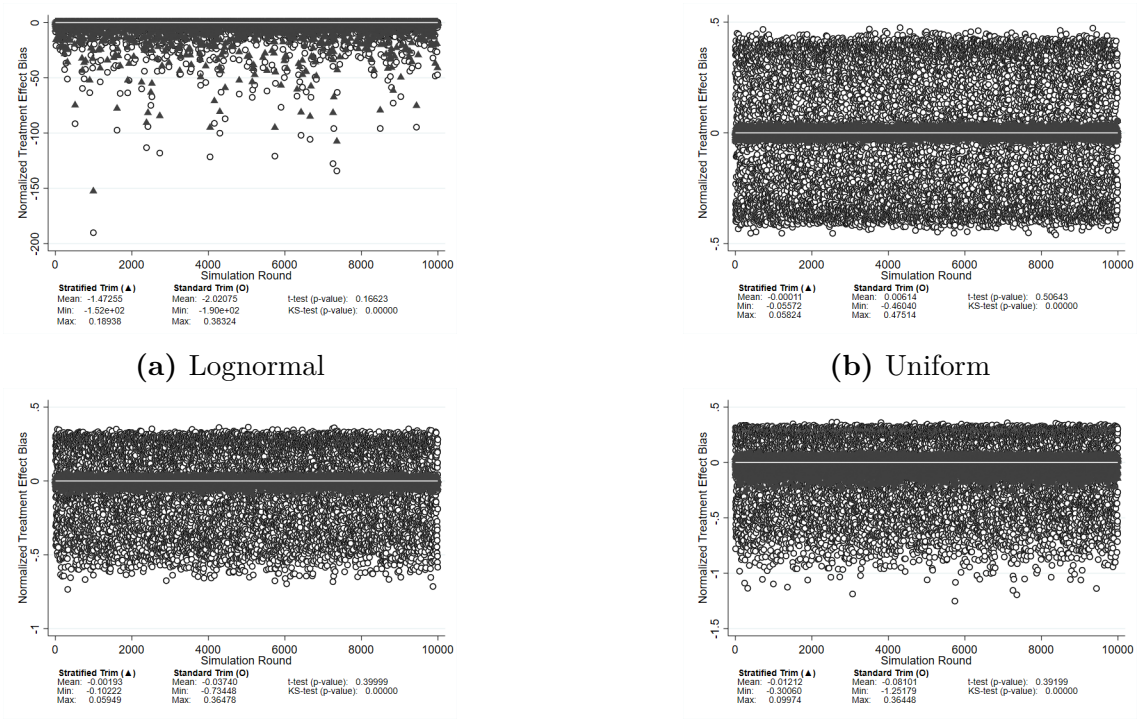
(e) Standard Deviation



(f) Skewness

Figure 9. Trimming: Normal Distribution

A.1.6 Trimming, Non-Normal Distribution



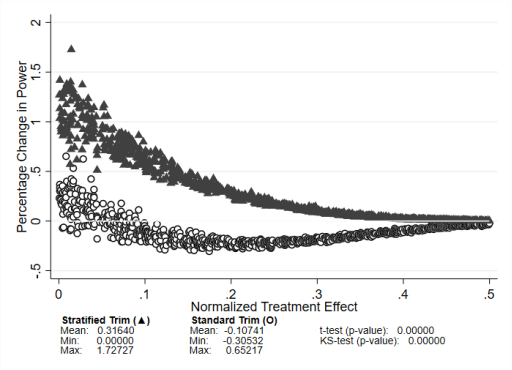
(a) Lognormal **(b)** Uniform

(c) Poisson **(d)** Gamma

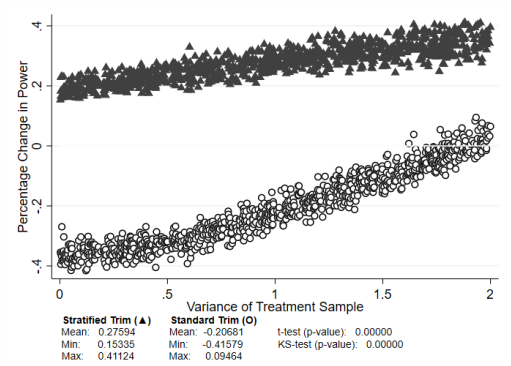
Figure 10. Trimming: Non-normal Distributions

A.2 Statistical Power

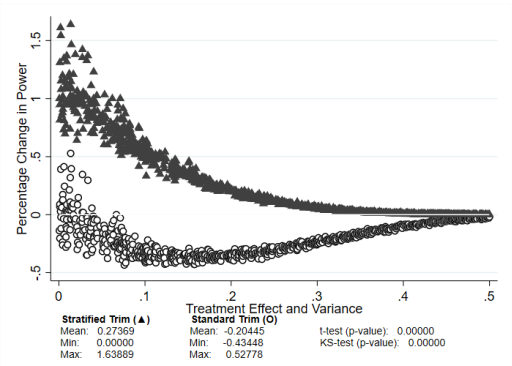
A.2.1 Trimming



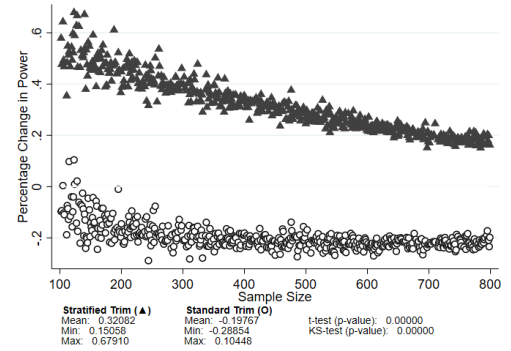
(a) Treat. Effect



(b) Variance



(c) Corr. Mean and Variance



(d) Sample Size

Figure 11. Effects of Trimming on Power Calculations

A.2.2 Percentage of Simulations with Improved Statistical Power

Tables 5 and 6 present the percentage of simulations in which the study's statistical power was reduced as a result of winsorizing and trimming - both the traditional and stratified approach - compared with no data manipulation. The findings suggest that it is very rare that *Stratified Winsorizing/Trimming* reduces a study's statistical power (less than 0.1% of cases), both compared with the traditional approach to winsorizing/trimming, and no winsorizing/trimming. On the other hand, the traditional approach to winsorizing and trimming frequently reduces a study's statistical power compared with no winsorizing or trimming, happening in 26% and 83% of cases, respectively.

Table 5: Percentage of Simulations with Improved Statistical Power

Winsorizing
A. % of simulations where *Strat. W/T* reduced statistical power compared with *Trad. W/T*
0.08
B. % of simulations where *Strat. W/T* reduced statistical power compared with *No W/T*
0.05
C. % of simulations where *Trad. W/T* reduced statistical power compared with *No W/T*
25.88

Notes: Standard errors are reported in brackets. Data is based on the each of the 1000 simulations underlying each sub-Figure of Figure 6, summed together.

Table 6: Percentage of Simulations with Improved Statistical Power

Trimming
A. % of simulations where *Strat. W/T* reduced statistical power compared with *Trad. W/T*
0.00
B. % of simulations where *Strat. W/T* reduced statistical power compared with *No W/T*
0.00
C. % of simulations where *Trad. W/T* reduced statistical power compared with *No W/T*
83.05

Notes: Data is based on the each of the 1000 simulations underlying each sub-Figure of Figure 11, summed together.

A.3 Type I Errors

Table 7: Frequency of Type I Errors

	Normal Distr.	Log-Normal Distr.	Skew-Normal Distr.	Gamma Distr.
B. Trimming				
No Trim	0.050	0.048	0.050	0.050
Traditional Trim.	0.050	0.050	0.050	0.050
Stratified Trim.	0.104	0.122	0.097	0.108
p-value No vs. Trad.	0.56	0.00	0.92	0.60
p-value No vs. Strat.	0.00	0.00	0.00	0.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00
B. Percentage of *No Trim* Type I errors included				
Traditional Trim.	38.13	23.05	42.12	34.80
Stratified Trim.	99.75	89.62	99.89	99.00
p-value Trad vs. Strat.	0.00	0.00	0.00	0.00

B Theory

B.1 Biased Treatment Effect

A researcher is interested in the relationship between Treatment T and outcome variable Y^*, where $T_i = \{0, 1\}$, with $T_i = 1$ is the treatment group, and $T_i = 0$ is the control group. However, Y_i^* contains white-noise measurement error η_Y, and hence the researcher only observes Y_i, where $Y_i = Y_i^* + \eta_Y$. The measurement error is uncorrelated with treatment status ($Cov(\eta_Y, T_i) = 0$). As such, the estimated regression can thus be written as $Y_i = \beta_1 T_i + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, which generates an unbiased estimate $\hat{\beta}_1$ of the true β_1 as $Cov(e_i, T_i) = 0$. Hence a white-noise measurement error in the outcome variable does not result in a biased estimate of the treatment effect.

B.1.1 Traditional Approach to Trimming

The researcher trims a share of the data, due to the fear that outliers are driving the estimates of β_1.[19] Hence the final outcome variable observed, and used by the researcher in their analysis, is $Y = Y^* + \eta_Y + \eta_T$, where η_T is the bias emerging as a result of trimming. The traditional approach to trimming can differentially trim the treatment and control groups. Therefore, $Cov(\eta_T, T_i) \neq 0$.

The estimated regression is thus: $Y_i = \beta_1 T_i + \eta_T + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, and the estimate $\hat{\beta}_1$ equals:

$$\hat{\beta}_1 = \frac{Cov(Y_i, T_i)}{Var(T_i)} = \beta \cdot \frac{Var(T_i)}{Var(T_i)} + \frac{Cov(\eta_T, T_i)}{Var(T_i)} + \frac{Cov(e_i, T_i)}{Var(T_i)} = \beta + \underbrace{\frac{Cov(\eta_T, T_i)}{Var(T_i)}}_{\text{bias} \neq 0}$$

Hence the use of trimming can result in a biased treatment effect estimate. This is because the trimming induced bias is correlated with the treatment assignment.

B.1.2 Stratified Trimming by Treatment

When *Stratified Trimming by Treatment* is used rather than trimming the entire sample, the final outcome variable observed and used by the researcher in their analysis is $Y_i = Y_i^* + \eta_Y + \eta_{STbT}$, where η_{STbT} is the bias as a result of *Stratified Trimming by Treatment*, with $Cov(\eta_{STbT}, T_i) = 0$.

[19] The conclusions are identical for winsorizing, expect that the selection bias is likely to be smaller as observations are not dropped, merely replaced.

The estimated regression is thus: $Y_i = \beta_1 T_i + \eta_{TbT} + \underbrace{\varepsilon_i + \eta_Y}_{e_i}$, and thus the estimate $\hat{\beta}_1$ equals:

$$\hat{\beta}_1 = \frac{Cov(Y_i, T_i)}{Var(T_i)} = \beta \cdot \frac{Var(T_i)}{Var(T_i)} + \frac{Cov(\eta_{STbT}, T_i)}{Var(T_i)} + \frac{Cov(e_i, T_i)}{Var(T_i)} = \beta$$

While $Cov(\eta_{STbT}, T_i) = 0$ is a strong assumption that does not necessarily always hold, so long as $Cov(\eta_{STbT}, T_i) < Cov(\eta_T, T_i)$, *Stratified Trimming by Treatment* will result in a lower bias on the treatment effect estimate ($\hat{\beta}_1$) than the traditional approach to trimming.

B.2 Type II Errors and Statistical Power

Trimming and winsorizing can also be used to improve a study's statistical power. By reducing the role of outliers, the variance of the distribution gets smaller, and hence statistical power increases. This is shown by the formula for the Minimum Detectable Effect (MDE), where a smaller value means higher power:

$$MDE = (t_{1-\kappa} + t_{\frac{\alpha}{2}}) \cdot \sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}} \tag{1}$$

The traditional approach to trimming improves statistical power by reducing the variance of the control and treatment group (σ_0^2 and σ_1^2), but also worsens power by decreasing each group's sample size (n_0 and n_1).[20] *Stratified Trimming by Treatment* can have differential effects on statistical power compared with the traditional approach to trimming, because it ensures that the tails of the distributions of both treatments are trimmed, while this is not guaranteed in the traditional approach to trimming. However, as illustrated in Figure 4, *Stratified Trimming by Treatment* can trim observations that are outliers of the Control or Treatment distributions, but are close to the mean of the entire sample. This can increase the variance of the sample's distribution. These two forces are counteracting, and hence it is unclear ex ante whether *Stratified Trimming by Treatment* improves or worsens a study's statistical power compared with the traditional approach to trimming.

Unlike the traditional approach to trimming, *Stratified Trimming by Treatment* ensures that proportionate shares of control and treatment groups are removed from the sample (see Section 3.1.1). Assuming the researchers divided the sample size across treatment and control to maximize statistical power (if there are two groups, then $P = 0.5$), *Stratified Trimming by Treatment* ensures the ratio between treatment and control is unchanged,

[20] Winsorizing does not result in a smaller sample size.

which improves statistical power.

B.2.1 Standard Errors

The standard error of $\hat{\beta}_1$ (in a regression $Y_i = \alpha + \beta_1 \cdot T_i + \varepsilon_i$) is given by the following formula:

$$SE(\hat{\beta}_1) = \sqrt{\frac{s^2}{Var(T_i)}} \tag{2}$$

where s^2 is the estimate of the error variance, given by $s^2 = \frac{RSS}{n-2}$, with RSS standing for the Residual Sum of Squares ($RSS = \sum(y_i - \hat{y}_i)^2$, with $\hat{y}_i$ being the predicted value of y for observation i). Measurement errors result in a larger Standard Error, as the RSS becomes larger due to the additional error term introduced (η_Y). Winsorizing/Trimming aims to reduce the RSS, resulting in a smaller Standard Error. Comparing the effects of the traditional approach to winsorizing/trimming versus *Stratified Winsorizing/Trimming by Treatment* does not highlight a clear winner. In some instances, *Stratified Winsorizing/Trimming by Treatment* will result in a smaller RSS, and in other instances a larger RSS. This is also supported by the applications to Augsburg et al. (2015), Angelucci et al. (2023), Schilbach (2019) and Jack et al. (2023) in Section 4 and Appendix D, where the standard errors in Panel A (the traditional approach to winsorizing/trimming) are sometimes larger, and other times smaller than the standard errors in Panels B and C (*Stratified Winsorizing/Trimming by Treatment* & *Treatment*TimePeriod*).

C Trimming Code

Researchers must decide along which dimension to stratify their trimming. In case of stratification along multiple variables (e.g., treatment status and survey round), a new variable needs to first be created that encodes this new strata.

C.1 Stata

Traditional approach to trimming:[21] winsor2 *OutcomeVar*, cuts(5 95) trim
Stratified Trimming: winsor2 *OutcomeVar*, cuts(5 95) by(*StratifiedVariable*) trim

C.2 R

Using the newly created package called WinsorByGroupR (see GitHub repository here):
Traditional approach to trimming: trim(data, value_col = "*OutcomeVar*", bounds = c(5, 95))
Stratified trimming: trim_by_group(data, group_col = "*StratifiedVariable*", value_col = "*OutcomeVar*", bounds = c(5, 95))

[21] This code trims *OutcomeVar* at the 5th and 95th percentile.

D Applications to Schilbach (2019) and Augsburg et al. (2015)

D.1 Schilbach (2019)

Schilbach (2019) conducted an RCT among cycle-rickshaw drivers in India, offering different monetary incentives for sobriety. The study included three groups: one receiving unconditional payments (*Control*), another receiving payments contingent on sobriety (*Incentive*), and a third group choosing their preferred incentive structure (*Choice*).

Table 8(i) replicates the findings in Table A.10 of Schilbach (2019) using the traditional winsorization technique (Panel A), and the *Stratified Winsorizing per Treatment* approach (Panel B). Both coefficients as well as their significance level increase as a result of the stratified approach to winsorization. Table 8(ii) captures the share of the three experimental conditions that are winsorized using both techniques. Compared with the traditional approach to winsorizing, *Stratified Winsorizing per Treatment* winsorizes less of the control group, and more of the *Choice* treatment arm.[22] As Table 8(ii) illustrates, *Stratified Winsorizing by Treatment* decreases the discrepancy in the percentage of observations winsorized across the three experimental arms. This impacts treatment effect estimates, however to a smaller extent than previous applications. As such, this presents a case where the two approaches to winsorizing illustrate the robustness of the treatment effect estimates to the chosen winsorizing technique.

D.2 Augsburg et al. (2015)

Augsburg et al. (2015) conduct an RCT in Bosnia and Herzegovina to evaluate the impact of microcredit loans, offered to loan applicants that were otherwise marginally rejected by a microfinance institution. The authors document an increase in profits, but no change in overall household income. To ensure outliers are not driving the results, the authors trim 1-3% of the right-tail of the outcome variables' distribution.

Table 9 reproduces Appendix Table A.10 from Augsburg et al. (2015), using both the traditional approach to trimming - the technique deployed by the authors - (Panel A), and *Stratified Trimming by Treatment* (Panel B). Table 9.(i) reports the OLS estimates of the estimated treatment effect of being offered a microfinance loan in Bosnia and Herzegovina,

[22] Schilbach (2019) winsorizes at both the left and right tail - however the winsorizing process only winsorizes the left tail of the distribution, as 6.39% of the respondents reported having the highest level of savings. Therefore, the intuition is the same as in Figure 3.

on the respondents' assets, business, and income.

Panel B is emphasized to indicate cases where the statistical significance of the treatment effect increased (in **bold**) and decreased (underlined) as a result of *Stratified Trimming by Treatment*, compared with the traditional approach to trimming. Overall, *Stratified Trimming by Treatment* improves the statistical significance of treatment effect estimates, however it can also reduce the statistical significance of estimates, particularly when 1% of the sample is trimmed. The interpretation of the effect of the microloan on business expenses, revenues, and profits does not change, but the treatment effect sizes increase by 77%, 87%, and 32%, respectively as a result of 3% *Stratified Trimming by Treatment*. The interpretation of the effect of microloans on income changes as a result of *Stratified Trimming by Treatment*: the treatment effect on welfare benefits is now statistically significantly negative, while it is a null result under the traditional approach to trimming.

To understand why *Stratified Trimming by Treatment* can improve or worsen an estimate's statistical significance, Table 9.(ii) reports the fraction of observations from treatment and control groups that are trimmed, separately for the traditional approach to trimming (Panel C), and *Stratified Trimming by Treatment* (Panel D). Again, these are emphasized in Panel D, with **bold** values representing cases in which *Stratified Trimming by Treatment* improved the statistical significance of the treatment effect, and underlined values representing cases where the statistical significance worsened as a result of *Stratified Trimming by Treatment*.

Columns (1) - (4) in Table 9 document regressions in which the Treatment has a positive treatment effect. Panel C illustrates that treatment and control group observations are trimmed disproportionately under the traditional approach to trimming. This discrepancy in the trimming of treatment and control observations is reduced in *Stratified Trimming by Treatment*, see Panel D. The cases in which *Stratified Trimming by Treatment* improves the statistical significance of treatment effects in Panel B (in **bold**) also correspond to the cases where *Stratified Trimming by Treatment* reduces the discrepancy between the share of trimmed right-tail observations from control and treatment, by decreasing the fraction of trimmed Treatment observations, increasing the fraction of trimmed control observations, or both (Table 9, Panel D). By trimming fewer right-tailed observations of the treatment group, the mean of the treatment group's distribution increases, and the difference between treatment and control increases. Similarly, by trimming more right-tailed observations of the control group, the mean of the control group's distribution decreases. The intuition is the same underlying Figure 3.

The cases in which *Stratified Trimming by Treatment* reduces the statistical significance

of treatment effect estimates in Table 9.(i) Panel B (underlined) are the cases where the traditional approach to trimming, "under-trims" the treatment group. *Stratified Trimming by Treatment* brings the fraction of trimmed observations in the treatment group closer to the fraction of trimmed observations in the control group.[23]

Columns (5) - (7) of Table 9.(i) report negative treatment effects of microfinance on household income. In this case, the effect is reversed: *Stratified Trimming by Treatment* increases the fraction of trimmed right-tail observations from the treatment group, or reduces the fraction of trimmed observations from the control group (Table 9.(ii), Panel D, Column (7)). By trimming more right-tailed observations of the treatment group, the mean of the treatment group's distribution decreases, and the difference between treatment and control increases. Similarly, by trimming fewer right-tailed observations of the control group, the mean of the control group's distribution increases.

The cases where the coefficient estimates in Panels A and B are the same (Column (6)) or don't change a statistically significant amount (Columns (1) and (5)) are due to standard errors being very large, or *Stratified Trimming by Treatment* not changing the share of trimmed observations that are from the treatment group substantially.[24]

[23] In the simulations, *Stratified Trimming* ensured the fraction of Treatment observations that were trimmed was always equal to the fraction of Control observations that were trimmed. With the data from existing papers, this is not always be the case, due to the structure of the data. For example, if the researcher wants to trim the lower 5%, however the bottom 10% of observations take the value of 0, no trimming will occur.

[24] The trimmed share of treatment and control in Column (6) is identical for the traditional trimming approach, and *Stratified Trimming by Treatment*. This is due to a spike in observations at the 99.55th percentile in the control group, and hence fewer observations get trimmed.

Table 8: Schilbach (2019), Table A.10

Table 8.(i) OLS Treatment Effect Estimates

	Dependent variable: Amount saved at study office (Rs./day)		
Fraction of winsorized data:	0%	1%	2%
	(1)	(2)	(3)
A. Traditional Winsorizing			
Incentives	11.28	13.43*	12.08*
	(6.22)	(5.42)	(5.13)
Choice	16.62**	16.19**	15.09**
	(5.58)	(5.17)	(4.97)
B. Stratified Winsorizing by Treatment			
Incentives	11.28	13.43*	12.16*
	(6.22)	(5.43)	(5.13)
Choice	16.62**	17.13**	16.65***
	(5.58)	(5.15)	(4.95)

Table 8.(ii) % of Treat. and Control Obs. Winsorized

Fraction of winsorized data:	0%	1%	2%
	(1)	(2)	(3)
C. Traditional Winsorizing			
% of Control Obs. Winsorized		0.57	1.01
% of Incentives Obs. Winsorized		0.44	0.59
% of Choice Obs. Winsorized		0.35	0.56
D. Stratified Winsorizing by Treatment			
% of Control Obs. Winsorized		0.44	0.89
% of Incentives Obs. Winsorized		0.44	0.59
% of Choice Obs. Winsorized		0.42	0.98

Notes: Standard errors are in parentheses. Stratified Winsorizing by Treatment winsorizes the sample separately for the three experimental arms, while Traditional Winsorizing winsorizes the entire sample. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01

Table 9: Augsburg et al. (2015), Table A.10

Table 9.(i) OLS Treatment Effect Estimates

	Assets & Business				Income		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Asset Value	Busi. Expenses	Busi. Revenues	Busi. Profits	Wages	Remittances	Benefits
A. Traditional Trimming							
1% Trim	2,265	552.7**	1,539**	858.9**	-235.5	-41.27	-94.58
	[6,326]	[249.8]	[639.1]	[405.3]	[446.3]	[84.73]	[64.61]
2% Trim	-2,451	323.4**	1,032**	896.7**	-236.6	-0.719	-54.03
	[5,878]	[159.2]	[470.7]	[351.2]	[409.6]	[68.38]	[58.61]
3% Trim	-414.5	260.8**	744.1*	648.0**	-346.7	18.85	-45.11
	[5,390]	[129.4]	[403.1]	[301.5]	[395]	[65.64]	[52.42]
B. Stratified Trimming by Treatment							
1% Trim	-3,963	467.1*	1,368**	672.3*	9.597	-41.27	**-140.4****
	[6,626]	[263.8]	[661.2]	[384.7]	[455.3]	[84.73]	[66.52]
2% Trim	-2,451	**548.3*****	**1,316*****	751.1**	-69.56	-0.719	**-135.0****
	[5,878]	[180.8]	[477.5]	[309.7]	[411.3]	[68.38]	[60.52]
3% Trim	-1,861	**462.9*****	**1,393*****	**853.4*****	-106.3	18.85	**-117.6****
	[5,464]	[134.5]	[434.3]	[284.4]	[397.2]	[65.64]	[55.18]

Table 9.(ii) % of Treatment and Control Observations Trimmed

	Assets & Business				Income		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Asset Value	Busi. Expenses	Busi. Revenues	Busi. Profits	Wages	Remittances	Benefits
C. Traditional Trimming							
C.1 1% Trim							
% of Control Obs. Trimmed	1.23	0.88	0.88	0.70	0.53	0.35	1.23
% of Treat. Obs. Trimmed	0.32	0.64	0.64	0.48	1.12	0.80	0.32
C.2 2% Trim							
% of Control Obs. Trimmed	1.41	1.23	1.23	1.06	1.23	1.23	2.11
% of Treat. Obs. Trimmed	1.59	1.75	1.75	0.80	2.07	1.28	0.64
C.3 3% Trim							
% of Control Obs. Trimmed	2.29	1.58	1.58	1.41	1.76	1.23	2.82
% of Treat. Obs. Trimmed	2.23	2.55	2.87	1.91	3.19	1.28	1.44
D. Stratified Trimming by Treatment							
D.1 1% Trim							
% of Control Obs. Trimmed	0.70	0.70	0.70	0.70	0.70	0.35	**0.70**
% of Treat. Obs. Trimmed	0.80	0.64	0.64	0.80	0.80	0.80	**0.64**
D.2 2% Trim							
% of Control Obs. Trimmed	1.41	**1.41**	**1.41**	1.41	1.41	1.06	**1.41**
% of Treat. Obs. Trimmed	1.59	**1.28**	**1.44**	1.59	1.59	2.18	**1.44**
D.3 3% Trim							
% of Control Obs. Trimmed	2.11	**2.11**	**1.94**	**2.11**	2.11	1.23	**2.11**
% of Treat. Obs. Trimmed	2.39	**1.91**	**1.75**	**1.91**	2.39	1.28	**2.23**

Notes: An 1% trim means the top 1 percentile of observations have been trimmed; similarly for the 2% and 3% trims. Standard errors are reported in brackets. Covariates included. Observation unit: respondent except income from self employment (household). BAM: Bosnia and Herzegovina convertible mark. The exchange rate at baseline was US$1 to BAM 1.634. Stratified Trimming by Treatment trims the sample separately for treatment and control, while Traditional Trimming trims the entire sample. Results are reported without corrections for multiple hypothesis testing. * p<0.1, ** p<0.05, *** p<0.01