Estimation of Average Total Effects in Quasi-Experimental Designs: Nonlinear Constraints in Structural Equation Models

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften
der Friedrich-Schiller-Universität Jena

von Dipl.-Psych. Joachim Ulf Kröhne

geboren am 02. Juni 1977 in Jena

Gutachter:

1. Prof. Dr. Rolf Steyer (Friedrich-Schiller-Universität Jena)
2. PD Dr. Matthias Reitzle (Friedrich-Schiller-Universität Jena)

Tag des Kolloquiums: 23. August 2010

Dedicated to

Cora and our family(ies)

Zusammenfassung

Diese Arbeit untersucht die Schätzung durchschnittlicher totaler Effekte zum Vergleich der Wirksamkeit von Behandlungen basierend auf quasi-experimentellen Designs. Dazu wird eine generalisierte Kovarianzanalyse zur Ermittlung kausaler Effekte auf Basis einer flexiblen Parametrisierung der Kovariaten-Treatment Regression betrachtet.

Ausgangspunkt für die Entwicklung der generalisierten Kovarianzanalyse bildet die allgemeine Theorie kausaler Effekte (Steyer, Partchev, Kröhne, Nagengast, & Fiege, in Druck). In dieser allgemeinen Theorie werden verschiedene kausale Effekte definiert und notwendige Annahmen zu ihrer Identifikation in nicht-randomisierten, quasi-experimentellen Designs eingeführt. Anhand eines empirischen Beispiels wird die generalisierte Kovarianzanalyse zu alternativen Adjustierungsverfahren in Beziehung gesetzt und insbesondere mit den Propensity Score basierten Analysetechniken verglichen. Es wird dargestellt, unter welchen Annahmen die generalisierte Kovarianzanalyse zu unverfälschten Schätzungen durchschnittlicher totaler Effekte führt und es wird ein Überblick über zusätzliche Herausforderungen gegeben, die für eine Analyse von durchschnittlichen totalen Effekten in nicht-randomisierten Designs berücksichtigt werden müssen.

Es werden drei zentrale Anforderungen für die Schätzung von durchschnittlichen totalen Effekten in quasi-experimentellen Designs mit Behandlungszuweisung auf individueller Ebene dargestellt: (1) Interaktionen zwischen Kovariaten und der Behandlungsvariablen, (2) die Stochastizität der Kovariaten sowie die daraus resultierende Nichtlinearität der zu prüfenden Hypothese sowie (3) Varianzheterogenität der Residuen.

Im theoretischen Teil dieser Arbeit wird gezeigt, dass der Standardfehler des allgemeinen linearen Modells für den durchschnittlichen totalen Effekt häufig verfälscht ist, wenn Interaktionen zwischen Kovariaten und der Behandlungsvariable vorliegen und die Kovariaten stochastische Größen sind. Ebenso wird gezeigt, dass auch für mittelwertszentrierte Kovariaten die Standardfehler für den durchschnittlichen totalen Effekt verfälscht sind, wenn die Kovariaten mit ihrem geschätzten Stichprobenmittelwert zentriert werden. Ausgehend von einer Darstellung der gebräuchlichen Methoden zur Erforschung von Interaktionstermen wird hergeleitet, dass für die statistische Inferenz über durchschnittliche totale Effekte die Annahme einer gemeinsamen multivariaten Verteilung der Erfolgsvariablen, der Kovariaten und der Behandlungsvariablen notwendig ist.

Die Annahme einer gemeinsamen Verteilung ist notwendig für unverfälschte Standardfehler von Schätzern durchschnittlicher totaler Effekte. Diese Annahme einer gemeinsamen Verteilung der Regressoren wird traditionell für Strukturgleichungsmodelle gemacht. Es wird deshalb untersucht, wie mit Hilfe von

Strukturgleichungsmodelle mit nichtlinearen Restriktionen geschätzter Modellparameter Hypothesen über durchschnittliche Behandlungseffekte getestet werden können und die bereits von Nagengast (2006) und Flory (2008) untersuchten Strukturgleichungsmodelle werden im Hinblick auf die drei oben genannten Anforderungen untersucht. Zusätzlich werden Verfahren der Standardfehlerkorrektur des allgemeinen linearen Modells für die statistische Inferenz über durchschnittliche totale Effekte erforscht, beispielsweise robuste Standardfehler basierend auf heteroskedastizitätskonsistenten Varianzkovarianzmatrizen (White, 1980a) sowie adjustierte Standardfehler für sogenannte *regression estimates* (Schafer & Kang, 2008). Spezifische Hypothesen über die Robustheit der Ansätze werden beschrieben. Insbesondere wird im Hinblick auf die theoretisch begründete Varianzheterogenität untersucht, unter welchen Bedingungen Robustheit für die jeweiligen Verfahren erwartet werden kann.

In zwei Monte-Carlo Simulationen werden die verschiedenen Strukturgleichungsmodelle mit nichtlinearen Restriktionen vertiefend analysiert und mit Prozeduren zur Standardfehlerkorrektur des allgemeinen linearen Modells verglichen. Dazu werden sieben konkrete Forschungsfragestellungen unter einem breiten Spektrum möglicher Parameterkonstellationen für die Datenerzeugung untersucht. Zusätzlich wird die Verfälschung des Standardfehlers bei Mittelwertszentrierung illustriert und es werden die Konsequenzen der Verletzung der Linearitätsannahme der allgemeinen linearen Hypothese aufgezeigt. Die Verfahren, welche in der ersten Simulation zu unverfälschten Schätzungen des durchschnittlichen totalen Effekts und zu unverfälschten Standardfehlern geführt haben, werden dann in einer zweiten Monte-Carlo Simulation im Hinblick auf die statistische Power und ihr Verhalten bei kleinen Stichprobengrößen verglichen.

Die Ergebnisse der Simulationsstudie bestätigen die Angemessenheit von richtig spezifizierten Strukturgleichungsmodellen mit nichtlinearen Restriktionen für die Analyse von durchschnittlichen totalen Effekten in Beobachtungsstudien. Bedeutsame Unterschiede zwischen Ein- und Mehrgruppenmodellen für die Schätzung von durchschnittlichen totalen Effekten werden gezeigt. Weiterhin wird der von Nagengast (2006) untersuchte und in *EffectLite* (Steyer & Partchev, 2008) implementierte Augmentierungsansatz der Varianz-Kovarianzmatrix der Parameterschätzer bestätigt. Die Simulationsstudie zeigt auch, dass der von Schafer & Kang (2008) entwickelte adjustierte Standardfehler für den durchschnittlichen totalen Effekt auch für stochastische Regressoren und unter allen betrachteten Bedingungen mit Varianzheterogenität sowie unabhängig von der Gruppengröße unverfälscht ist. Der direkte Vergleich der Strukturgleichungsmodelle mit nichtlinearen Restriktionen mit Schafer & Kangs adjustiertem Standardfehler demonstriert die Angemessenheit der generalisierten Kovarianzanalyse als Strukturgleichungsmodell mit nichtlinearen Restriktionen. Die statistische Power der Wald-Test Statistik der nichtlinearen Restriktionen ist unter den analysierten Datensätzen der zweiten Simulationsstudie insgesamt für mittlere und große Stichprobengrößen

vergleichbar mit der statistischen Power der Teststatistik basierend auf dem adjustierten Standardfehler für *regression estimates*.

Abschließend werden die verschiedenen Vor- und Nachteile der generalisierten Kovarianzanalyse als erweitertes Mehrgruppen-Strukturgleichungsmodell vergleichend zu dem adjustierten Standardfehler für *regression estimates* dargestellt. Den spezifischen Vorteilen, (1) latente Kovariaten berücksichtigen zu können, (2) eine flexible Behandlung fehlender Werte insbesondere der Kovariaten zu ermöglichen und (3) eine Erweiterbarkeit für den Vergleich von mehr als zwei Behandlungsgruppen zu ermöglichen, steht vor allem die multivariate Normalverteilungsannahme als Nachteil gegenüber, welche für die untersuchten Ansätze zur Schätzung der Parameter der Strukturgleichungsmodelle notwendig ist.

In der Diskussion wird die Bedeutung der Unterscheidung zwischen fixierten und stochastischen Regressoren für die Analyse durchschnittlicher totaler Effekte hervorgehoben. Darüber hinaus werden weiterführende Forschungsfragen und mögliche Ergänzungen der untersuchten Analysetechniken, bspw. im Hinblick auf eine Kombination mit Propensity Score basierten Adjustierungsverfahren, dargestellt. Die Arbeit schließt mit konkreten Empfehlungen für die Anwendung der generalisierten Kovarianzanalyse und für die Weiterentwicklung des Programms zur Analyse kausaler Effekte (*EffectLite*, Steyer & Partchev, 2008).

Abstract

This thesis focuses on estimating average total effects for the comparison of treatments based on quasi-experimental designs. For this purpose a generalization of analysis of covariance will be considered for the estimation of causal effects based on a flexible parameterization of the covariate-treatment regression.

The stochastic theory of causal effects (Steyer, Partchev, Kröhne, Nagengast, & Fiege, in press) constitutes the starting point for developing generalized analysis of covariance. In this general theory, different causal effects are defined and sufficient assumptions for their identification in non-randomized, quasi-experimental designs will be introduced. Using an empirical example, we will compare generalized analysis of covariance to the various other adjustment techniques and, in particular, we will compare this method to adjustment procedures based on propensity scores. Furthermore, we treat some assumptions implying unbiased estimates of average total effects. We will also discuss additional challenges in applying adjustment procedures for causal effects in non-randomized designs.

We will describe three issues that are crucial for the estimation of average total effects in quasi-experimental designs with treatment assignment on the individual level: (1) Interactions between covariates and the treatment variable, (2) stochasticity of covariates and the resulting nonlinearity of the hypothesis of no average total effect, as well as, (3) heterogeneity of residual variances.

In the theoretical part of this thesis we will show that the standard error for the average total effect estimator of the general linear model is often biased if there are interactions between covariates and the treatment variable and if the covariates are stochastic. Accordingly, hypotheses about average total effects cannot be tested within the general linear model if the covariates are stochastic. Furthermore, we will show that the standard error of the average total effect will also be biased for mean-centered covariates, if the covariates are centered on their estimated sample means. Based on a review of the traditional methods for probing interaction terms we will point out that assuming a joint distribution of the outcome variable, the treatment variable and the covariates is necessary for valid statistical inference about average total effects.

Assuming a joint distribution is necessary for developing unbiased standard errors for estimators of the average total effect. This assumption of a joint distribution of regressors is traditionally made in structural equation models. Therefore, we will study how structural equation models can be used for testing hypotheses about average total effects via nonlinear constraints of estimated parameters, and procedures already used by Nagengast (2006) and by Flory (2008) will be analyzed in detail with respect to the three requirements mentioned above. Additionally, we will investigate different methods of standard error correc-

tion in the general linear model with respect to statistical inference about average total effects, for instance, robust standard errors based on heteroscedasticity-consistent estimators of the variance-covariance matrix (White, 1980a) and adjusted standard errors for the so-called regression estimates (Schafer & Kang, 2008). We will describe specific hypothesis about the robustness of the approaches. In particular, we will study under which conditions robustness with respect to heterogeneity of residual variances can be expected for the different approaches.

In two Monte Carlo simulations we will analyze different structural equation models with nonlinear constraints and compare them to various corrected standard errors of the general linear model. We will investigate seven concrete research questions which can be answered in a simulation study. In the first simulation study the different procedures are investigated under a broad range of possible parameter constellations. Additionally, we will illustrate the bias of the estimators standard error of the average total effect for mean-centered covariates. Similarly, the consequences of the violated linearity assumption of the general linear hypothesis are studied in some detail. Those methods yielding unbiased estimates of the average total effect and unbiased standard errors in the first simulation study will be compared with respect to power and small sample behavior in the second part of the Monte Carlo simulation.

The results of the simulation study confirm the adequacy of appropriately specified structural equation models with nonlinear constraints for the analysis of average total effects in observational studies. It is shown that there are important differences between single group and multi-group models for the estimation of average total effects. Furthermore, the augmentation approach of the variance-covariance-matrix of parameter estimates already used by Nagengast (2006) and implemented in *EffectLite* (Steyer & Partchev, 2008) will be confirmed. Our simulation study also shows that the adjusted standard errors for the average total effect developed by Schafer and Kang (2008) are unbiased under all considered conditions with stochastic covariates and heterogeneous residual variances regardless of group sizes. The appropriateness of generalized analysis of covariance as a structural equation model with nonlinear constraints is demonstrated by a direct comparison to Schafer and Kang's adjusted standard errors. The statistical power of our Wald-test statistic for the nonlinear constraint is comparable overall to the tests based on the adjusted standard errors of the least-squares regression estimates for all simulated data sets of the second Monte Carlo simulation with medium and large sample sizes.

Finally, different advantages and disadvantages of generalized analysis of covariance as an enhanced multi-group structural equation model are described comparing it to the adjusted standard errors for regression estimates. The specific benefits, (1) the incorporation of latent covariates, (2) the flexibility to deal with missing values, in particular, of the covariates and, (3) the extendibility to multi-group comparisons

on the one hand are contrasted by the necessary assumption of multivariate normality used to estimate the parameters of the structural equation model on the other hand.

In the discussion we highlight the importance of the distinction between fixed and random regressors for the analysis of average total effects. Moreover, subsequent research questions and further extensions of the adjustment methods are described, for instance, the combination of generalized analysis of covariance with propensity score based approaches. The thesis closes with practical recommendations for the application of generalized analysis of covariance and further developments of the program package for the analysis of causal effects (*EffectLite*, Steyer & Partchev, 2008).

Acknowledgements

I owe many people thanks for completing my thesis and it is a pleasure to list at least the names of the many people who made this thesis possible. In the beginning I would like to thank my first supervisor Rolf Steyer for sharing his expertise with me, for giving me the opportunity to work at his department as well as for the support, challenges and scientific surrounding he provided for his PhD-students. Matthias Reitzle as my second supervisor encouraged me through all the years to keep focused on my research interests and provided help in so many ways. The especially fruitful surrounding was made up by colleagues, fellow PhD-students and members of Rolf's staff. In particular I would like to thank Ivailo Partchev for his wonderful way of being a truly statistician who taught me a lot. For the friendly and supportive atmosphere, for professional and technical advices and for the emotional support I thank all my fellow PhD-students and colleagues: Andreas Wolf, Felix Flory, Benjamin Nagengast, Steffi Pohl, Hendryk Böhme, Jan Marten Ihme, Tim Loßnitzer, Christiane Fiege, Norman Rose, Anja Vetterlein, Erik Sengewald, Sonja Hahn, Axel Mayer and Uwe Altmann. For those of you who are not ready yet: I wish you all the best for the completion of your own dissertation. In addition, Katrin Schaller and Marcel Bauer warrant recognition concerning their help in administrative, technical and organizational needs. Moreover, representative thanks go to Christof Nachtigall and Marc Müller form the project *kompetenztest.de* for many years of cooperative work and for giving me the opportunity to work practically as well. This thesis also highly profited from the yearly colloquium on research methods in empirical educational research in Riezlern / Hirschegg. Representative for all the participants over the last years I especially want to thank the organizers Johannes Hartig and Andreas Frey for the possibility of sharing ideas and having great discussions. Apart from my supervisors and colleagues, I owe special gratitude to my friends for continuous and unconditional support of all my undertakings. Finally, I would love to thank all my family, in particular my father and his partner.

Contents

List of Figures

List of Tables

Listings

Chapter 1
Introduction

This thesis is concerned with causal inference in observational studies – more precisely, with statistical inference about average total effects estimated in quasi-experimental between-group designs.

The issue of the *causal inference* can be described in the following elementary way: Consider a specific treatment that can be present (indicated by $X = 1$) or absent (indicated by $X = 0$) for an observational unit $U = u$. The primary interest of causal inference is to assess if there is a difference in a subsequently measured outcome variable Y that is due to the difference in the two treatment conditions (present or absent) for a population of observational units. For our purposes, the definition of a meaningful difference between the observed outcomes of treated and untreated units on the population level needs special attention because, for instance, for between-group designs, only one of the two different treatment conditions $X = 1$ or $X = 0$ is realizable for each unit u .

Sir Ronald A. Fisher (1925/1946) introduced the term *randomization* for a specific design technique that makes causal inference possible whenever the treatment assignment is completely under the control of the researcher for each unit u (see, e. g., Krauth, 2000, for historical remarks). If the units can be assigned randomly to one of the treatment conditions (or if the treated units are randomly selected), and if the randomization does not fail, we can identify and estimate the *average total effect* of the treatment on the outcome variable Y under very general conditions. If the observational units are sampled randomly from the population, the sample means of the subsequently measured outcome variable Y can be computed for the treated and the untreated units. The difference between both means yields an estimate of the average total effect that is due to the difference in the two treatment conditions. Under these circumstances *statistical inference* (e. g., the hypothesis that the average total effect of the treatment is zero) is possible using the simple t–test for independent groups.

For *quasi-experimental designs*,[1] where the assignment to different treatment conditions is, by definition, not randomized (see, e. g., Shadish, Cook, & Campbell, 2002), the simple mean difference is known to be misleading (see, e. g., Campbell & Stanley, 1963). Nevertheless, although the average total effect is not identified as a simple mean difference, an average total effect that is due to the different treatment conditions is still defined and can be identified and estimated under specific circumstances. We will summarize

[1] We use the term quasi-experimental designs instead of the more general term *observational studies* (see, e. g., Wold, 1956; Cochran & Chambers, 1965; McKinlay, 1975; Rosenbaum, 2002a, 2010) to show that although the treatment is not randomly assigned, e. g., due to self-selection or investigator selection, the applications we are interested in share the common purpose to test hypotheses on *manipulable causes*.

the necessary definitions as well as the fundamental assumptions for the identification of average total effects in a probabilistic theoretical framework from a *pre-facto perspective* in the next section. Based upon this theoretical framework, we will then introduce the *generalized analysis of covariance* as an adjustment method described by Steyer, Partchev, Kröhne, Nagengast, and Fiege (in press) for the analysis of average total effects in quasi-experimental designs. Finally, in the third section of this introduction, we will sum up how the specific research question that is studied within this thesis — the implementation of this generalized analysis of covariance as a structural equation model — will be approached. In the second chapter, we will provide an empirical example for the estimation of average total effects from quasi-experimental studies and illustrate different data analysis techniques for the estimation of adjusted average total effects.

1.1 Theory of Causal Effects

The general theory of stochastic causality provided by Steyer et al. (in press) serves as the theoretical background for the definition of the average total effect and for the derivation of basic requirements of the statistical model used to test hypotheses about this average total effect. The basic components necessary for the introduction of the theoretical foundation of the average total effects for designs without randomization are the outcomes under each treatment condition. These outcomes usually cannot be observed simultaneously because, for instance, for between-group designs, each unit can only be observed in one treatment condition at a time. This was called the fundamental problem of causal inference (Holland, 1986, see also H. I. Weisberg, 1979, for an early illustration of the fundamental problem of causal inference). Although we cannot observe the outcomes under different treatment conditions simultaneously, the *true outcome variables* can be defined for each of the treatment conditions. These true outcome variables are introduced in the subsequent subsection.

1.1.1 True Outcome Variables

The elementary components for considering causal effects are defined either in a stochastic way (for instance, as *true-yields* used by Neyman, 1923/1990, or as the expectations of the outcome variable Y given the treatment conditions $X = 1$ or $X = 0$ and the unit $U = u$ used by Steyer, Gabler, von Davier, & Nachtigall, 2000; Steyer, Gabler, von Davier, Nachtigall, & Buhl, 2000; Steyer, Nachtigall, Wüthrich-Martone, & Kraus, 2002), or in a deterministic way (like the so-called *potential outcomes* $Y_{X=j}(u)$ and $Y_{X=k}(u)$, see, e. g., Rubin, 1974, 1977, 1978, as the basic references for the *Rubin Causal Model*). As described by Steyer et al. (in press), the *deterministic outcome assumption*, i. e., the assumption $Var(Y|X,U) = 0$ underlying the potential outcomes, is a conceptual limitation which is too restrictive for a general theory. Moreover, in order to define meaningful theoretical quantities that are purged from confounding, conditioning on the unit $U = u$ (the

smallest possible subpopulation) is not sufficient for a general theory at the *most fine-grained level*. Accordingly, the definition of the *true outcome variables* must take into account *pre-* and *equi-orderedness* of random variables, to also capture, for instance, confounding on an individual level, and to include mediating processes on a theoretical level. Therefore, the *true outcome variables* for each treatment condition j of X — as used in this thesis in line with Steyer et al. (in press) — are obtained by conditioning on all potential confounders, and not only on the observational-unit variable U. This is technically carried out by defining the true outcomes variable as (versions of) the following regressions

$$\tau_j \equiv E_{X=j}(Y|\mathfrak{D}_X), \tag{1.1}$$

where $\mathfrak{D}_X$ denotes a *confounder* σ*-algebra* of X, constructed in such a way that all events and variables that are pre- or equi-ordered (i. e., simultaneous) to X are measurable w.r.t. $\mathfrak{D}_X$. The values of τ_j are the conditional expectations of Y in an *atomic stratum* (i. e., given combinations of values on all potential confounders), and given the treatment group $X = j$.

A covariate represents an attribute of the unit *prior* to or *at* the onset of a treatment. Hence, a covariate can never be affected by the treatment, and all potential confounders which are prior or simultaneous to X can be considered as covariates. We use the term covariate for the subset $Z = (Z_1, ..., Z_K)$ of potential confounders which is used in data analysis for the estimation of causal effects. Nevertheless, conditioning on $\mathfrak{D}_X$ in the definition of the true outcome variable means controlling for all potential confounders. Note that the conditional expectation in Equation (1.1) is a regression with respect to a conditional probability measure $P_{X=j}$, i. e., it is only uniquely defined conditional on $X = j$. A complete definition of the true outcome variable is omitted here because it is not guaranteed that the true outcomes — i. e., the values of a true outcome variable — are uniquely defined for different treatment conditions without further assumptions (see Steyer et al., in press, ch. 5, for mathematical details). In the following we will assume that the true outcome variable τ_j is uniquely defined for each treatment $X = j$, and that it is *unconfounded* due to the conditioning on the confounder σ-algebra $\mathfrak{D}_X$ that comprises all covariates (and all potential confounders).

The idea of defining the core part of the theory of stochastic causality, the true outcome variables, by conditioning the outcome variable on all potential confounders makes the theory flexible enough to handle problems that are not covered by the traditional Rubin Causal Model (Rubin, 1974, 1977, 1978), for instance confounding on the individual level (see also Sobel, 1995, p. 20).

1.1.2 True and Conditional Total Effects

Subsequent to the definition of the true outcome variables themselves, we are now introducing the *true total effect variable*. The true total effect variable δ_{jk} is defined as the difference between the true outcome variable τ_j for treatment group $X = j$ and the true outcome variable τ_k for the comparison group $X = k$. This difference can be interpreted as the *unconfounded* effect of the treatment $X = j$ (for instance, $X = 1$) compared to the treatment $X = k$ (for instance, $X = 0$), i. e., the values of the true effect variable δ_{jk} are the true total effects in the most fine-grained strata of the potential confounders:

$$\delta_{jk} \equiv \tau_j - \tau_k. \tag{1.2}$$

The conditional expectation of the true effect variable δ_{jk} with respect to a value v of the random variable V is called *conditional total effect* [$CTE_{jk;V=v} \equiv E(\delta_{jk}|V = v)$], and the regression of the true effect variable δ_{jk} on V is called *conditional total effect function* [$E(\delta_{jk}|V)$, also labeled as *conditional total effect variable*]. The conditional total effect $E(\delta_{jk}|U = u)$ for the comparison of treatment $X = j$ and $X = k$ for a given person $U = u$ is of particular interest, also labeled as the *individual total effect* ($ITE_{jk;U=u}$):

$$ITE_{jk;U=u} \equiv CTE_{jk;U=u} = E(\delta_{jk}|U = u). \tag{1.3}$$

The difference of the true outcomes is equal to the *individual total effects* $E_{X=j}(Y|U = u) - E_{X=k}(Y|U = u)$ for the special case of $\mathfrak{D}_X = \sigma(U)$, i. e., the values of δ_{jk} in Equation (1.2) are equal to the difference between the two conditional expectations of Y given an observational unit $U = u$ and given the experimental conditions $X = j$ or (respectively) $X = k$ (under the assumption that $\tau_j = E_{X=j}(Y|U)$ is uniquely defined, see above).

Neither the true total effect nor the individual total effect (or the conditional total effect given a person) is identifiable in general for a unit $U = u$. It is impossible to observe the same unit in more than one treatment condition (while everything else is kept constant). This means that although the *true effect variables* are constructed in such a way as to be purged from confounding, it is necessary to consider the expectations of the true total effect variable over a distribution or a conditional distribution (of potential confounders).

1.1.3 Average Total Effects

In order to identify a net causal effect for the comparison of two treatments in between-group designs, we are now focusing on the *average total effect,* ATE (also known as *average treatment effect*), defined in the

theory of stochastic causality as the expectation of the true total effect variable over the distribution of the atomic strata, which is equal to the difference between the expectations of the true outcome variables:

$$ATE_{jk} \equiv E(\delta_{jk}) = E(\tau_j) - E(\tau_k). \tag{1.4}$$

The (unconditional) average total effect is also equal to the expectation of conditional total effect variables for a variable V, i. e., $E(CTE_{jk;V}) = E\big(E(\delta_{jk}|V)\big) = E(\delta_{jk}) = ATE_{jk}$ (see Steyer et al., in press, for details, and the substantive meaning for different choices of the variable V), and because the individual total effect variable is a special conditional total effect variable, the average total effect equals the expectation of the individual total effect variable $ITE_{jk;U} \equiv E(\delta_{jk}|U)$ as well, i. e., $E(ITE_{jk;U}) = ATE_{jk}$. Hence, this average total effect is presented as a generalization of the average treatment effect as defined in the Rubin Causal Model based on the deterministic potential outcomes $Y_{X=j}(u)$ and $Y_{X=k}(u)$ [Steyer et al., in press].[2]

1.1.4 Prima Facie Effects and Biases

After defining the average total effect, we will now briefly discuss conditions under which this effect can be identified with empirically estimable parameters. Therefore, let us start with the definition of the *prima facie effect* PFE_{jk} of treatment $X = j$ compared to treatment $X = k$,

$$PFE_{jk} \equiv E(Y|X = j) - E(Y|X = k), \tag{1.5}$$

i. e., as the difference between the conditional (i. e., group-specific) expectations of the outcome variable Y in treatment group $X = j$ and in treatment group $X = k$ (Holland, 1986). This quantity can be estimated, for instance, as the difference between the observed means of the outcome variable Y between the groups $X = j$ and $X = k$ under common random sampling conditions. Throughout this thesis we will call the regression $E(Y|X)$ of Y on X *treatment regression*.

In order to describe under which conditions the *PFE*, i. e., the "at-first-sight-effect", can be interpreted as causal effect of the treatment, Steyer et al. (in press) consider the following decomposition of the prima facie effect into the average total effect of interest and two bias components (see also Morgan & Winship, 2007):

$$PFE_{jk} = ATE_{jk} + \textit{baseline bias}_{jk} + \textit{effect bias}_{jk}. \tag{1.6}$$

[2] Note that beside the average total effect ATE_{jk}, i. e., the (unconditional) expectation of the true effect variable, various conditional expectations given a value v of a variable V are considered in the theory of stochastic causality, for instance, conditional total effects given a value of the covariate, $E(\delta_{jk}|Z = z)$, and the conditional total effects given a value of the treatment variable, $E(\delta_{jk}|X = x)$ for $x = j$ or $x = k$. The conditional expectation of the true total effect variable over the conditional distribution in the treatment group $X = j$ is known as *average treatment effect of the treated, ATT*. Although the average total effects given a treatment condition can be estimated using generalized analysis of covariance studied in this thesis, we restrict the presentation to the average total effect ATE_{jk} (see, e. g., Steyer & Partchev, 2008, for details).

The first bias component related to the true outcome τ_k for the comparison group, i. e., the *baseline bias*$_{jk} = E(\tau_k|X = k) - E(\tau_k|X = j)$, is zero if the treatment probabilities are independent from the true outcomes in group k (i. e., there is no selection due to the expected outcomes in the treatment condition $X = k$). The second bias component related to the true effect, i. e., the *effect bias*$_{jk} = E(\delta_{jk}|X = j) - ATE_{jk}$, is zero if the $(X = j)$-conditional expectation of the true effect variable for the treatment group does not deviate from the average total effect (i. e., there is no selection due to the expected effect of the treatment condition $X = k$ compared to the treatment condition $X = j$). If these bias components, the baseline bias and the effect bias, are both equal to zero or cancel out each other, the average total effect equals the prima facie effect, that is, the average total effect equals the true mean difference.

Furthermore, *conditional prima facie effects*, e. g., prima facie effects conditional on a value $Z = z$ of a covariate, can be defined as

$$PFE_{jk;Z=z} \equiv E(Y|X = j, Z = z) - E(Y|X = k, Z = z). \tag{1.7}$$

The values of the regression $E(Y|X, Z)$ are the conditional expectations $E(Y|X = j, Z = z)$ of the outcome variable Y given treatment $X = j$ and the value $Z = z$ of the covariate. We will use the term *covariate-treatment regression* for this regression $E(Y|X, Z)$ of Y on X and Z, where the covariate Z might be univariate or multivariate $Z = (Z_1, ..., Z_K)$.

The corresponding conditional *PFE*-functions $PFE_{jk;Z} \equiv E_{X=j}(Y|Z) - E_{X=k}(Y|Z)$ can be decomposed, similar to the unconditional prima facie effect, as follows:

$$PFE_{jk;Z} = CTE_{jk;Z} + \textit{baseline bias}_{jk;Z} + \textit{effect bias}_{jk;Z}. \tag{1.8}$$

1.1.5 Unbiasedness

The two bias components introduced in the last subsection can be used to formulate the condition of *unbiasedness* of the prima facie effect, which is according to Equation (1.6) equivalent to the expression that the two bias components cancel each other out, i. e.,

$$\textit{baseline bias}_{jk} = -\textit{effect bias}_{jk}, \tag{1.9}$$

and, as a special case, the PFE_{jk} is also unbiased if both bias components are zero. In other words, the prima facie effect is unbiased if $PFE_{jk} = ATE_{jk}$ [see Equation (1.6)].

Furthermore, the treatment regression is called unbiased, if

$$E(Y|X=j) = E(\tau_j), \quad \text{for each value } j \text{ of } X, \tag{1.10}$$

i. e., $E(Y|X)$ is unbiased if the expectation of the true outcome variable τ_j equals the $(X=j)$-conditional expectation of the outcome variable Y for all treatment groups j. Unbiasedness of the treatment regression implies that the PFE_{jk} is unbiased and can be used to identify the average total effect.

In addition to unconditional unbiasedness of the treatment regression, the covariate-treatment regression $E(Y|X,Z)$ is defined to be Z-conditionally unbiased, if for each treatment group j of X the $(X=j)$-conditional regression of the outcome variable Y on the covariate Z is equal to the regression of the true outcome variable τ_j on the covariate Z (and if the true outcome variable τ_j exists, see Steyer et al., in press):

$$E_{X=j}(Y|Z) = E(\tau_j|Z) \text{ almost surely (}a.s.\text{)} \quad \text{for each value } j \text{ of } X. \tag{1.11}$$

If the covariate-treatment regression is Z-conditionally unbiased, the conditional *PFE*-functions are equivalent to the conditional total effect functions $CTE_{jk;Z}$. Therefore, the values of the *PFE*-functions can be used to identify exactly these conditional total effect functions and accordingly the average total effect can be identified as $E(PFE_{jk;Z}) = E(CTE_{jk;Z}) = ATE_{jk}$.

1.1.6 Causality Conditions

Unbiasedness of the treatment regression and Z-conditional unbiasedness of the covariate-treatment regression are the weakest criteria presented by Steyer et al. (in press) to obtain average total effects based on the empirically estimable regressions $E(Y|X)$ or $E(Y|X,Z)$. The central assumption for the discussion of generalized analysis of covariance as structural equation model with nonlinear constraints is Z-conditional unbiasedness of the covariate-treatment regression. Therefore, we will summarize the sufficient conditions implying unbiasedness in this subsection.

Independent Cause For designs with *randomization* it is well known that the simple observed mean difference is an estimator of the (average) total effect. In terms of the theory of stochastic causality, the expectations $E(\tau_j)$ of the true outcome variables can be identified by the conditional (group-specific) expectations $E(Y|X=j)$ of the outcome variable if the treatment assignment is randomized. This follows from the fact that perfect randomization[3] will ensure independence of X and a confounder σ-algebra $\mathfrak{D}_X$, i. e., independence of the treatment variable X and all potential confounders, abbreviated as $X \perp\!\!\!\perp \mathfrak{D}_X$. Independence of X and $\mathfrak{D}_X$ is the first causality condition that implies unbiasedness of $E(Y|X)$. Stochastic

[3] For a discussion of randomization in small samples see, e. g., Hsu (1989) or Altman (1998).

independence of X and $\mathfrak{D}_X$ is equivalent to the following statement about the conditional treatment probability $\phi_j \equiv P(X = j|\mathfrak{D}_X)$ and the unconditional treatment probability $P(X = j)$ if X represents a treatment variable:

$$P(X = j|\mathfrak{D}_X) = P(X = j) \;\; a.s. \text{ for each value } j \text{ of } X. \tag{1.12}$$

In a similar way *conditional randomization* or *randomization based on a covariate* (see, e. g., Rubin, 1977; H. I. Weisberg, 1979; Sobel, 1998; Steyer et al., 2002) will ensure conditional independence of X and $\mathfrak{D}_X$ given a (possibly multivariate) covariate Z, denoted as $X \perp\!\!\!\perp \mathfrak{D}_X|Z$. Conditional independence, which, for a treatment variable X with $J+1$ discrete values, is equivalent to

$$P(X = j|\mathfrak{D}_X) = P(X = j|Z) \;\; a.s. \text{ for each value } j \text{ of } X, \tag{1.13}$$

implies Z-conditional unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ with respect to total effects.

As long as the randomization is based on the *observed value* of the (possibly multivariate) covariate, conditional randomization will ensure conditional independence no matter whether the covariate is measured without error, i. e., $Z = f(U)$, or if the covariate is a fallible measure with error, i. e., $Z = f(U) + \varepsilon$.

Complete Cause Conditions A further pair of causality conditions is discussed by Steyer et al. (in press): (Unconditional) *Completeness* of the regression $E(Y|X)$ and *Z-conditional completeness* of the regression $E(Y|X,Z)$. The idea of completeness of the regression $E(Y|X)$ is that none of the potential confounders affects the expectations of Y over and above the treatment variable X. Accordingly, the treatment regression $E(Y|X)$ is called *complete* ($Y \vdash \mathfrak{D}_X|X$) if

$$E(Y|X,\mathfrak{D}_X) = E(Y|X) \;\; a.s., \tag{1.14}$$

and the covariate-treatment regression $E(Y|X,Z)$ is called *Z-conditionally complete* ($Y \vdash \mathfrak{D}_X|X,Z$) if

$$E(Y|X,\mathfrak{D}_X) = E(Y|X,Z) \;\; a.s. \;. \tag{1.15}$$

The condition $Y \models \mathfrak{D}_X|X,Z$ means that for the covariate-treatment regression $E(Y|X,Z)$ the conditional expectation of the outcome Y remains unaffected by any additional (potential) confounder in $\mathfrak{D}_X$ over and above the possibly multivariate Z and the treatment variable X.

Completeness implies unbiasedness of $E(Y|X)$ and Z-conditional completeness implies Z-conditional unbiasedness of $E(Y|X,Z)$.

Strong Ignorability A third causality condition — the independence and conditional independence of X and the true outcomes — is mentioned here because it is well known in the literature of the Rubin Causal Model. Under the assumption that τ_j exists for each j of X and with $\tau \equiv (\tau_0, ..., \tau_J)$ as the vector of true outcomes, independence of X and the true outcomes ($\tau \perp\!\!\!\perp X$) can be formulated as

$$P(X = j|\tau) = P(X = j) \;\; a.s. \text{ for each value } j \text{ of } X. \tag{1.16}$$

Conditional independence of X and true outcomes ($\tau \perp\!\!\!\perp X|Z$), i. e.,

$$P(X = j|Z, \tau) = P(X = j|Z) \;\; a.s. \text{ for each value } j \text{ of } X, \tag{1.17}$$

is a probabilistic formulation of Rubin's *"strong ignorability"*. Note that the additional requirement of the Rubin Causal Model, that $0 < P(X = j|Z) < 1$ (see Rosenbaum & Rubin, 1983b), is replaced by the assumption that the true outcome variable exists (or by requiring $P(X = j|Z) \underset{a.s.}{>} 0$, see Steyer et al., in press, for details). The strong ignorability assumption is also known as the assumption of *no unmeasured confounders* (see, e. g., Tsiatis, 2006, ch. 13).

Z-conditional independence of X and the true outcomes imply that the covariate-treatment regression $E(Y|X, Z)$ is unbiased. Furthermore, for the true treatment probability functions $\phi_j \equiv P(X = j|\mathfrak{Q}_X)$ (also called the *true propensity functions*), it follows in the same way that the regression $E(Y|X, \phi)$ with $\phi = (\phi_1, ..., \phi_J)$ is unbiased if

$$P(X = j|\phi_j, \tau) = P(X = j|\phi_j) \;\; a.s. \text{ for each value } j = 1, ...J \tag{1.18}$$

(see Steyer et al., in press, for the proof). If X and the true outcomes are Z-conditionally independent, the Z-conditional propensity functions $\pi_j \equiv P(X = j|Z)$ can be used for the estimation of the average total effects instead of the true propensity function ϕ_j. If Z is a (possibly K–dimensional, multivariate) covariate, the vector $\pi = (\pi_1, ..., \pi_J)$ of Z-conditional propensity functions fulfills the requirements of a covariate as well. Therefore, the dimensionality of the covariates can be reduced from K [the number of covariates in $Z \equiv (Z_1, ..., Z_K)$] to J (the number of groups minus one, i. e., to one for the simple comparison of two treatment groups).

Z-conditional independence of X and the true outcomes implies unbiasedness of the regression $E(Y|X, \pi)$.

Unconfoundedness All conditions mentioned so far imply that the baseline as well as the effect bias are both equal to zero. *Unconfoundedness* — as introduced by Steyer, Gabler, von Davier, and Nachtigall (2000)

— is the weakest falsifiable condition for unbiasedness, which only implies that baseline and effect bias cancel each other out. One possibility to define unconfoundedness is to consider *potential confounders*, which are by definition all random variables measurable with respect to $\mathfrak{D}_X$. The treatment regression $E(Y|X)$ is called unconfounded, if for each value j of X and for each *potential confounder* W either

$$P(X = j|W) = P(X = j), \quad \text{or} \tag{1.19}$$

$$E_{X=j}(Y|W) = E_{X=j}(Y) \tag{1.20}$$

is (almost surely) true. Two additional equivalent definitions, formal proofs of the implications of unconfoundedness as well as a full length discussion of the implication structure between the causality conditions mentioned in this introduction are omitted here as these can be found in detail in Steyer et al. (in press). In the following, we restrict our attention to the usability of unconfoundedness as the theoretical basis for the selection of covariates as summarized in the next subsection.

1.1.7 Selection of Covariates

If unbiasedness of the treatment regression $E(Y|X)$ cannot be assumed, the well-known strategy is to consider the *covariate-treatment regression* $E(Y|X,Z)$. This strategy is called *conditioning on covariates* in various research traditions (see, e. g., Steyer et al., 2002, and also Heckman & Vytlacil, 2007). As summarized by Steyer, Fiege, and Rose (2010) , *"selecting the appropriate covariates is the only option in quasi-experimental studies"* to identify (average) total effects, i. e., the selection of a set of $Z \equiv (Z_1, ..., Z_K)$ covariates for which one of the sufficient conditions (e. g., *Z-conditional independence*, $X \perp\!\!\!\perp \mathfrak{D}_X|Z$, or *$Z$-conditional completeness*, $Y \vdash \mathfrak{D}_X|X,Z$) hold is essential, for instance, for estimating average total effects.

It is often argued that the selection of covariates requires substantive subject matter knowledge (see, e. g., Hernán, Hernández-Díaz, Werler, & MitcheII, 2002, and references therein for a discussion in epidemiology). Also T. D. Cook, Shadish, and Wong (2008) and Pohl, Steiner, Eisermann, Soellner, and Cook (2009) highlight the importance of a careful covariate selection based on theoretical considerations. As the authors point out, the choice of covariates is — compared to the decision between different adjustment methods — tremendously important for the estimation of unbiased average total effects.

As summarized by Schafer and Kang (2008), *"introducing more covariates reduces the residual dependence of X on τ_0 and τ_1, which helps to eliminate the selection bias arising from nonrandom treatment assignment"* (Schafer & Kang, 2008, p. 282, notation changed). However, even if the covariate-treatment regression is unbiased, the inclusion of an additional covariate might actually re-introduce bias (see, e. g., Pearl, 2000, 2003, 2009). Nevertheless, when many (continuous or qualitative) covariates are considered,

the application of one of the adjustment procedures (see subsection 2) might become difficult (Rosenbaum, 2002b, p. 76). Variable-selection methods (e. g., stepwise regression with respect to the mean squared error) are not appropriate for obtaining unbiasedness (Schafer & Kang, 2008). Therefore, a criterion based on the theory of stochastic causality useable for covariate selection is of great interest.

The inclusion of a potential confounder W as an additional covariate should be considered if the covariate-treatment regression is not Z-conditionally unconfounded for a given set of covariates Z (Steyer et al., in press). This is by definition the case if, for the potential confounder W, none of the following two conditions for Z-conditional unconfoundedness hold for a value $Z = z$ and a group j:

$$P_{Z=z}(X = j|W) = P(X = j|Z = z), \quad \text{or} \tag{1.21}$$

$$E_{X=j,Z=z}(Y|W) = E_{X=j}(Y|Z = z). \tag{1.22}$$

If neither Equation (1.21) nor Equation (1.22) is true for a pair of (j,z), the potential confounder W should be included in the vector of covariates $Z = (Z_1, ..., Z_K)$ for which we postulate Z-conditional unbiasedness.

Steyer et al. (in press) developed a second strategy to falsify unfoundedness of a covariate-treatment regression with respect to a potential confounder W. For this purpose, they showed that the covariate-treatment regression is Z-conditionally unconfounded if and only if for each potential confounder W the following statement is true:

$$E_{X=j}(Y|Z) = E\left(E_{X=j}(Y|Z,W)|Z\right). \tag{1.23}$$

For discrete covariates Z and for a discrete potential confounder W (as well as for a treatment variable X with $J+1$ discrete values as considered in this thesis) Equation (1.23) can be easily checked. Nevertheless, for continuous and probably multivariate covariates Z and / or for a continuous potential confounder W, assumptions about the functional form might be necessary to model the $(X = j)$-conditional regressions of Y on Z and W as well as to estimate their unconditional expectation in Equation (1.23). We will discuss functional form assumptions for the covariate-treatment regression in more detail in the next subsection. Here it is important to highlight that a functional form assumption might be necessary not only for the estimation of (adjusted) average total effects (see next section) but also for the falsification of (Z-conditional) unconfoundedness with respect to a potential confounder W.

1.1.8 Adjusting and Functional Form Assumption

In this subsection we will summarize how adjusted average total effects can be estimated based on the assumption of Z-conditional unbiasedness and under which circumstances an additional functional form

assumption is necessary, even though Z-conditional unbiasedness holds because it is implied by one of the mentioned causality conditions with respect to a (possible multivariate set of) covariate(s) Z.

For quasi-experimental designs without randomization, the assumption of Z-conditional unbiasedness of the covariate-treatment regression (see subsection 1.1.5) provides the theoretical justification for all adjustment methods considered in this thesis, either based on covariate adjusted means or based on Z-conditional propensities. The following two subsections can be seen as the heart of the theory with respect to the identification and parameterization of average total effect estimators. Covariate adjusted means will be used for generalized analysis of covariance, and Z-conditional propensities can be seen as an alternative for the identification of an average total effect.

Covariate Adjusted Means Unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ implies that the expectation $E(\tau_j)$ of the true outcomes can be computed from $(X = j)$-conditional regressions $E_{X=j}(Y|Z)$ of the outcome variable Y on the covariate(s) Z as follows:

$$E(\tau_j) = E\big(E_{X=j}(Y|Z)\big). \tag{1.24}$$

Estimates of the expectations $E(\tau_j)$ in Equation (1.24) are called *adjusted means* of the outcome variable Y (see, e. g., Milliken & Johnson, 2002). According to the definition of the average total effect [see Equation (1.4)], the difference between the true adjusted means for two treatment groups j and k equals the average total effect , i. e., $ATE_{jk} = E\big(E_{X=j}(Y|Z)\big) - E\big(E_{X=k}(Y|Z)\big)$.

If all covariates Z are discrete, no further assumptions are necessary in order to model the $(X = j)$-conditional regression in Equation (1.24). This is possible by specifying an ANOVA-like cell-mean model (also called *fully flexible coding*, e. g., Morgan & Winship, 2007, or saturated model, e. g., Angrist & Pischke, 2009), where the number of parameters equals the number of observed patterns of the (qualitative) covariate values. We will call this approach *fully saturated modeling* and remark that these saturated regression models are inherently linear in the dummy indicators used for the cell-mean model.

For the cases of continuous covariates or for discrete covariates with many distinct values, fully saturated modeling might be impossible for finite samples. Therefore, the specification of a non-statured regression model[4] is one possibility to identify the adjusted means based on the estimated parameters of a covariate-treatment regression for empirical applications (see, e. g., Sobel, 1998). We will refer to adjustment methods based on the covariate-treatment regression as *outcome modeling* when conditional on $X = j$ a functional form assumption is made *only* for the regression of Y on Z (also labeled as y-model, see,

[4]Note that in contrast to the term *nonparametric* regression, non-statured regression models are often classified as *parametric regressions* (see, e. g., Ruppert, Wand, & Carroll, 2003, ch. 2).

e. g., Kang & Schafer, 2007a). In addition to the assumption of Z-conditional unbiasedness, the specification of the covariate-treatment regressions functional form is an additional model assumption which can be misspecified in empirical applications.

Z-Conditional Propensities If the treatment assignment and the true outcomes are conditionally independent given Z, the Z-conditional propensities can be used to identify the expectations of the true outcome variables as well, i. e.,

$$E(\tau_j) = E\left(E_{X=j}(Y|\pi)\right). \tag{1.25}$$

Unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ implies that the conditional total effects $CTE_{jk;Z=z}$ are identified by the conditional prima facie effects $PFE_{jk;Z=z}$ given the value of the covariate $Z = z$ [see Equation (1.7)], and that the average total effect can be obtained by taking the expectation of the conditional *PFE*s over the distribution of the covariate(s). The same is true for the conditional total effects $CTE_{jk;\pi=p}$ given the value p of the Z-conditional propensity provided that Z-conditional independence of X and the true outcomes holds. Although the conditional total effects $CTE_{jk;\pi=p}$ given the value p of the Z-conditional propensity might be less informative than the conditional causal effects $CTE_{jk;Z=z}$ given the value of the covariate $Z = z$, the average total effect can be obtained as the expectation of the prima facie effects $PFE_{jk;\pi=p}$ given the value p of the Z-conditional propensity π, i. e., $ATE_{jk} = E\left(PFE_{jk;\pi}\right)$.

If the Z-conditional propensities are not known in quasi-experiments, it is necessary to estimate them. Similar to the estimation of the adjusted means $E_{X=j}(Y|Z)$ based on a fully saturated regression, the Z-conditional propensities can be estimated without further assumptions regarding a functional form for discrete covariates Z. However, fully saturated modeling of the treatment assignment might again be impossible for continuous covariates or for discrete covariates with many distinct values. Under these circumstances the estimation of the propensity scores can be performed based on a non-saturated model of the treatment probabilities, i. e., $P(X = 1|Z)$. Similar to the outcome modeling approach, the functional form assumption used for the estimation of the Z-conditional propensities can be misspecified in empirical applications (see, e. g., Shaikh, Simonsen, Vytlacil, & Yildiz, 2009). We will refer to approaches which are based on a functional form assumption *at least* for $P(X = 1|Z)$ as methods with *assignment modeling*. Note, however, that for propensity score based adjustment methods an additional outcome model is formulated for the estimation of average total effects.

Unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ also implies that the adjusted means can be computed by using weighted outcome variables Y_W. The weights are computed from the (known) Z-conditional propensities, or based on the estimated Z-conditional propensities (see section 2.3.2 and

Steyer et al., in press, for details). Again, if the true Z-conditional propensities are unknown, assignment modeling can be applied to estimate them under an additional functional form assumption.

Notwithstanding the assumption of Z-conditional unbiasedness of the covariate-treatment regression holds, the bias of the estimated average total effect depends on the correct specification of the particular regression model if non-saturated regression models are applied either for outcome modeling or for assignment modeling (see, e. g., Tan, 2006, for a similar argumentation). If the functional form is not modeled correctly, estimates of the average total effect may be biased (see section 2.5 for a discussion of the robustness of adjustment methods against the misspecification of the functional form assumption), even though all relevant potential confounders are included in the vector of covariates $Z = (Z_1, ..., Z_K)$.

1.1.9 Single-Unit Trial and Sampling

Single-Unit Trial The theory of stochastic causality as summarized thus far is formulated by Steyer et al. (in press) in terms of probabilistic concepts, i. e., events, random variables, (conditional) expectations and regressions. All random variables mentioned in this introduction refer to a random experiment, the so-called *single-unit trial*, to which the stochastic dependencies between these variables refer. For the simple quasi-experimental designs upon which this thesis focuses, the single-unit trial consists of:

a) Sampling an observational unit u from a population of units,

b) (Assessing the values z of the covariate(s) Z),

c) Assigning the unit or observing its assignment to one of several treatment conditions $X = j$ and

d) Recording the value y of the outcome variable Y.

The second step (b) of this random experiment is written in parenthesesbecause this part is only necessary if at least one of the covariates in Z is not a deterministic function of the unit variable U, i. e., $Z \neq f(U)$. For instance, covariates measured with error, i. e., *fallible covariates*, are not deterministic attributes of the observational unit (their values are not fixed given the sampled unit $U = u$). Therefore, the assessment of the covariates' values is considered as an additional step in the single-unit trial.

Various single-unit trials are discussed by Steyer et al. (in press).The single-unit trials are sufficient for the *definition* of all concepts and quantities of interest as well as to derive conditions under which we can identify the theoretical quantities by empirically estimable quantities.

Sampling and Population Model A sample is necessary to apply the theory of stochastic causality for the *estimation* of average total effects, as we will illustrate in the next chapter. We assume that this sample is the result of independent and identical replications of the described single-unit trial, an assumption that implies that we obtain a *simple random sample* with mutually independent observations.

To clarify the distinction between the single-unit trials and the sampling model, Steyer et al. (in press) and Nagengast (2009) describe a simple random experiment of tossing a fair coin as an example: The probability $P(H)$ of heads is well-defined prior to the experiment [and if the coin is fair the probability is known to be $P(H) = 0.5$], even if the coin is never flipped. No sample is necessary for the *definition* of the probability of heads. However, a sample is needed for the *estimation* of this probability from the observed relative frequency of heads. Obviously, we have to assume that the distributions and parameters which characterize the single-unit trial do not change between replications. In other words, this means again that we have to define a single-unit trial which is the empirical phenomenon we are studying.

To link this sampling procedure to the assumptions of the different adjustment methods as well as to the data generation for the Monte Carlo simulation (see chapter 4), we additionally refer to a model described by Schochet (2009) for randomized experiments, which assumes that the observations are a random sample (*super-population* model). The values of the units' true outcomes for $X = 1$ and $X = 0$ are described as draws from two distributions of true outcomes in the super-population, with finite means and variances [discussed by Schochet, 2009, for the special case with the restriction $\mathfrak{D}_X = \sigma(U)$]. The difference between the distributions of true outcomes (i. e., the distribution of the conditional causal effect variable $CTE_{10;U}$, see subsection 1.1.2) defines the distribution of the individual causal effects in the super-population.

Stable Unit Treatment Value Assumption Finally, we have to mention that in the tradition of the Rubin Causal Model the additional *stable unit treatment value assumption* (SUTVA) is routinely made.[5] According to Rubin (1990), the two most common ways in which SUTVA can be violated are *"(a) there are versions of each treatment varying in effectiveness or (b) there exists interference between units."* The first violation is captured by the theory of stochastic causality by defining only one indistinguishable treatment variable in the assignment process of the single-unit trial. Hence, if different versions of a treatment exist (for example different therapists conducting special kinds of a therapy), and if these differences between the treatments are relevant to be distinguished, an extended single-unit trial is necessary. The second typical violation of the SUTVA assumption becomes clear from a formulation of the assumption presented by Rubin (1990) as *"...the implicit assumption that the value $Y_{ij}(t)$ that would be observed for the j-th outcome on unit i if all units were exposed to treatment t is stable in the sense that it would take the same value for all other treatment allocations such that unit i receives treatment t."* This independence of the true outcome of a unit $U = u$ from the treatment assignment of other units is likely to be violated, for example, in multi-level

[5] Rubin (1986, p. 961) described this assumption in the following way: *"SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive."*

designs where the treatment is applied to groups (see, e. g., Nagengast, 2009). If the independence of the true outcomes in the mentioned sense cannot be assumed, an extended single-unit trial is again necessary.

For the review of adjustment methods presented in the next chapter and for the discussion of generalized analysis of covariance, we assume that the treatment is assigned on the individual level and that the single-unit trial, as formulated in this subsection, describes appropriately the considered empirical phenomenon underlying the quasi-experimental between-group design upon which the average total effect is estimated. Possible extensions and further research questions, for example, regarding the SUTVA assumption for multi-level designs are summarized in section 5.4.2. Nevertheless, even the formulation of the simple single-unit trial described above clarifies that the random variables U, Z, X, and Y as defined above have a *joint distribution* defined on a common probability space.

1.2 Generalized Analysis of Covariance

Steyer et al. (in press) introduce a technique for the analysis of data obtained from quasi-experimental designs called *generalized analysis of covariance*. Generalized analysis of covariance can be considered as an *outcome modeling* approach under the assumption of a Z-conditional unbiasedness of the covariate-treatment regression. The average total effect is estimated as the difference between the adjusted means introduced in subsection 1.1.8.

Decomposition The covariate-treatment regression can be written for generalized analysis of covariance as

$$E(Y|X,Z) = g_0(Z) + g_1(Z)\cdot I_{X=1} + ... + g_J(Z)\cdot I_{X=J}, \tag{1.26}$$

where $g_0(Z)$ is the *intercept function* and $g_j(Z)$ are the the $j = 1,...,J$ *effect functions*. The variables $I_{X=j}$ indicate (with the values 1 and 0) the membership of a unit $U = u$ in treatment group $X = j$. Each effect function $g_j(Z)$ is equal to the differences $E_{X=j}(Y|Z) - E_{X=0}(Y|Z)$. According to Equation (1.26), the average total effect comparing the treatment group $X = j$ to the comparison group $X = 0$ equals the expectation of the effect function $g_j(Z)$, that is

$$ATE_{j0} = E\big(g_j(Z)\big). \tag{1.27}$$

The decomposition in Equation (1.26), called the *fundamental equation for generalized analysis of covariance* (Steyer et al., in press), is *always* true and does not impose any restrictions. A similar decomposition is presented, for example, by Wooldridge (2001) based on mean-centered covariates.

Statistical Inference As mentioned before, for randomized trials the hypothesis that the average total effect of a treatment $X = j$ compared to treatment $X = 0$ is zero can be tested with the simple t–test for

independent groups. A similar hypothesis of no average total effect for the case of two groups and based on generalized analysis of covariance under the assumption of a Z–conditional unbiased covariate-treatment regression for non-randomized quasi-experimental designs is:

$$H_0 : ATE_{j0} = E\left(E_{X=j}(Y|Z) - E_{X=0}(Y|Z)\right) = E\left(g_j(Z)\right) = 0. \tag{1.28}$$

As we have already discussed in subsection 1.1.8, a functional form assumption is often necessary for continuous covariates when the sample size is limited. We will study different approaches to test the *no average total effect* hypothesis for linear parameterizations of the intercept function $g_0(Z)$ and the effect functions $g_j(Z)$.[6] Without restricting generality, we will focus on the comparison of two groups [i. e., on the intercept function $g_0(Z)$ and the effect function $g_1(Z)$]:

$$\begin{aligned} E(Y|X,Z) &= g_0(Z) + g_1(Z)\cdot X \\ &= \left(\gamma_{00} + \gamma_{01}\cdot Z\right) + \left(\gamma_{10} + \gamma_{11}\cdot Z\right)\cdot X. \end{aligned} \tag{1.29}$$

Note that the regression coefficient γ_{11} for the predictor $Z\cdot X$ in Equation (1.29) represents the covariate-treatment interaction. Covariate-treatment interactions are often described as the case of *non-parallel* regression lines (Rogosa, 1980) in the tradition of the analysis of covariance. In this tradition the covariate-treatment regression is formulated based on group-specific regressions $E_{X=j}(Y|Z) = \beta_{j0} + \beta_{j1}\cdot Z$, which fit into the structure in the following way: $E(Y|X,Z) = \left(\beta_{00} + \beta_{01}\cdot Z\right) + \left((\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})\cdot Z\right)\cdot X$. According to parameterization in Equation (1.29), the hypothesis of no average total effect can be formulated as

$$H_0 : ATE_{10} = E\left(g_1(Z)\right) = E\left(\gamma_{10} + \gamma_{11}\cdot Z\right) = \gamma_{10} + \gamma_{11}E(Z) = 0. \tag{1.30}$$

Based on the reviewed theory we will discuss different approaches to test this hypothesis, for instance, with the help of the general linear hypothesis and based on structural equation models with nonlinear constraints.

1.3 Summary and Outlook

In the first chapter of this thesis, we introduced the theoretical background for defining an average total effect for quasi-experimental designs without randomization. Based on the theory of stochastic causality we showed that, provided the covariate-treatment regression is unbiased, the average total effect can

[6] Note that linear parameterizations of the intercept and the effect function include interaction between covariates (e. g., $Z_3 \equiv Z_1\cdot Z_2$) as well as, for instance, quadratic terms (e. g., $Z_4 \equiv Z_2^2$) because these terms may be defined as additional covariates in the vector of multivariate covariates $Z = (Z_1, ..., Z_K)$ [see Steyer et al., in press].

be identified in quasi-experimental designs as the difference between the adjusted means. Moreover, the theory of stochastic causality provides causality conditions which imply unbiasedness of the covariate-treatment regression and which can be falsified in empirical applications. We also discussed conditions under which, in addition to the selection of all relevant covariates, a functional form assumption is necessary. According to this functional form assumption, adjustment methods can be classified as outcome modeling (when a non-saturated covariate-treatment regression is specified), and as assignment modeling (when the Z–conditional propensities are estimated by a non-saturated regression model). An underlying random experiment, the single-unit trial, was formulated which is the empirical phenomenon we consider in quasi-experiments. Finally, we summarized the fundamental distinction of intercept and effect function for generalized analysis of covariance in the last subsection. Even for the simple linear parameterized intercept and effect function, generalized analysis of covariance incorporates covariate-treatment interactions.

Main effects (i. e., average effects) are sometimes not considered in analysis of covariance when interactions are present (see, e. g., Myers & Well, 2003). Although the theoretical foundation for the estimation of average effects from covariate-treatment regressions with interaction terms was summarized in subsection 1.1.8, we will provide a second justification of the appropriateness of average total effects by comparing generalized analysis of covariance to other adjustment methods in chapter 2. This presentation will also serve to justify the deliberate choice of the statistical models used to derive an estimator with valid properties (e. g., consistency, unbiasedness and known asymptotic variances or known distribution) that are necessary to derive appropriate test statistics for the average total effect. Accordingly, we will focus especially on the issue of statistical inference about the estimated average total effect in the review of different adjustment procedures given in the next chapter. Furthermore, we will discuss practical issues concerning the application of different adjustment strategies to highlight both similarities as well as important differences between the data analysis techniques. Finally, an empirical example will illustrate some similarities between different adjustment methods which were originally developed from different theoretical perspectives and which are sometimes discussed controversially in the literature.

In chapter 3, we will focus exclusively on statistical issues for the implementation of generalized analysis of covariance. Three requirements for the statistical model shall be deduced from the theory of stochastic causality: The need to model covariate-treatment interactions, the necessity to incorporate heterogenous residual variances, and a commonly ignored statement about the stochastic nature of the covariates and the treatment variable. In light of these requirements, we will discuss the general linear model and the common treatment of interaction terms in the analysis of covariance tradition. In particular, we will discuss the (unconditional) variance of the average total effect estimator under the usual assumptions for ordinary least-squares regressions. This will enable us to describe and to identify the conditions under which the

general linear model is not suited for statistical inference about the estimated average total effect. We will then turn to structural equation models and discuss the implementation of generalized analysis of covariance in this framework. Finally, we will summarize research questions a) for a comparison of the different developed structural equation models and b) for the robustness of the different approaches based on ordinary least-squares regressions. These research questions will be studied in chapter 4 with two simulation studies. An overall interpretation of the results of these simulation studies as well as general conclusions are presented in chapter 5.

Chapter 2
Adjustment in Quasi-Experimental Designs

Subsequent to the introduction of the theory of stochastic causality, we will now give a survey of important data analysis procedures that may be used for estimating and testing conditional and average total effects in quasi-experimental designs. We will discuss the approaches as possible alternatives to the structural equation models with nonlinear constraints studied in this thesis and relate them to an empirical quasi-experimental example.

As indicated by Schafer and Kang (2008, p. 280), *"even under the assumption of unconfoundedness, causal inference is not trivial; many solutions have been proposed, and there is no consensus among statisticians about which methods are best."* Based on the assumption that the covariate-treatment regression $E(Y|X,Z)$ is Z-conditionally unbiased, data analysis can be done in different ways. The majority of approaches can be classified as outcome modeling on the one hand, or as assignment modeling on the other. An exhaustive discussion of adjustment methods and their technical details is beyond the scope of this section (see, e. g., Shadish et al., 2002; Gelman & Meng, 2004; Gelman & Hill, 2007; Morgan & Winship, 2007; Guo & Fraser, 2010, for a broad collection of methods). The selection of techniques presented here in more detail is motivated by the concrete empirical application. We present this quasi-experimental study in the subsequent subsection to show that even for a simple linear parameterized covariate-treatment regression, unbiased estimated average total effects can be obtained.[1]

2.1 Example for a Quasi-Experimental Treatment Evaluation

As an illustration of how to estimate average total effects from quasi-experimental data by different adjustment methods, imagine the evaluation of two group training programs: A mathematical training designed to improve students' abilities in math and a language training to enhance the English skills of students. To adjudicate on the usefulness of these intervention programs, we want to conduct an evaluation study. In order to do so, we try to find participants for an empirical comparison, for example, undergraduate students from the introductory courses in statistics at our university. After sampling the students (at random), we measure their ability in maths and English (as pre-test scores) and ask them to fill out a questionnaire. Finally, because randomized assignment of students to treatment conditions is not possible due to ethical or organizational considerations, we offer them the choice between one of the two training courses. In the following phase, the students receive the training according to their choice, applied individually and with-

[1] Note that this conclusion can be drawn only because of special properties of the example, described in detail in subsection 2.6.

out interference among the students. Finally, we measure the performance of each student as the outcome variable. Again, each student is measured in each of the two domains, regardless of the individual selected treatment condition. Hence, the simple idea of our evaluation is that students treated in one subject area should outperform untreated students, i. e., students with the mathematical training should perform better in the math-related outcome measured after the intervention than students who underwent the language training. With the same idea in mind, we also try to evaluate the effect of the language training as we might expect that the students treated in the English treatment condition outperform the comparison group.

A study of this kind was conducted by Shadish, Clark, and Steiner (2008a), discussed by R. J. Little, Long, and Lin (2008); Hill (2008); Rubin (2008a); Shadish, Clark, and Steiner (2008b), and also replicated by Pohl, Steiner, Eisermann, Soellner, and Cook (submitted). We shall discuss these studies in more detail here for two reasons: On the one hand, the authors have applied a majority of the different adjustment methods reviewed in the next subsection. We will therefore refer to these studies as illustrations of the plain formulas of the different strategies. On the other hand, the quasi-experimental evaluation of the two training programs was integrated in an additional randomized experiment, as we will explain later. Therefore, these studies are examples of the so-called *within-study-comparisons*. Within-study-comparisons are part of a broader class of research designs, in which random assignment and self-selection are combined (also known as *hybrid intervention trials*, see Qi, Little, & Lin, 2008), and are also discussed as *doubly randomized preference trials (DRPT)* in this research tradition. A similar two-arm design was also suggested by Rücker (1989) to separate self-selection and treatment preference effects (see also Zelen, 1990).

Due to the combination of random assignment and self-selection, these designs can be applied to judge the appropriateness of the applied adjustment methods. For this purpose, the adjusted average total effect from the quasi-experimental study described so far is compared to the estimated average total effect obtained from an additional randomized trial (under the assumption that no treatment preference effect is present). The complete design of the within-study comparison of Shadish et al. (2008a) is presented in Figure 2.1. For the review of adjustment methods, we shall concentrate on the right arm of the design (the non-randomized, quasi-experiment) and the estimation of the following two causal effects:

- The average total effect of the *language training* compared to the mathematics training on the post-test in the language domain and,
- The average total effect of the *mathematics training* compared to the language training on the mathematics outcome measured as post-test.

Both (adjusted) average total effects can be estimated from the non-randomized experiment (group C and group D in Figure 2.1, where the students choose one of the two training conditions, self-selection)

using one of the data analysis techniques described in the following subsections provided that the necessary assumptions summarized in chapter 1 are fulfilled.

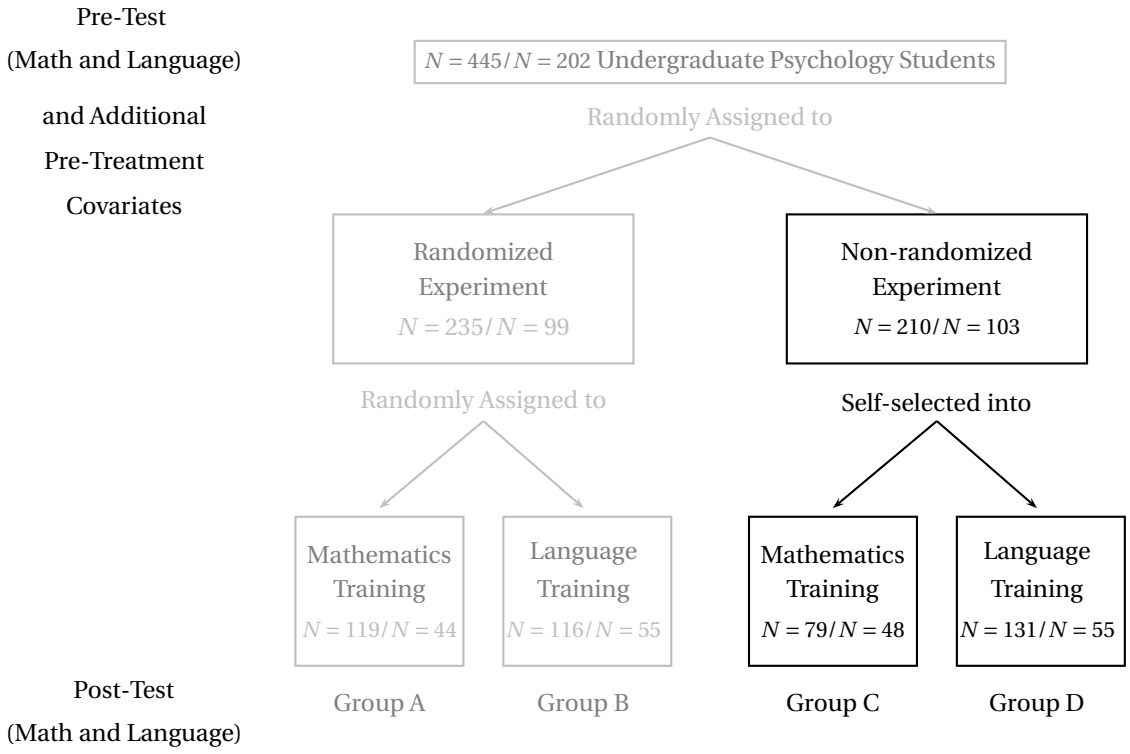


Figure 2.1: Design of the evaluation study used as introductory example

In subsection 2.6, we will again pick up the randomized part of the example and unveil the promising methodological benefit from the comparison of group A and group B (left arm of the design in Figure 2.1 printed in gray). As we will show in the light of the results of this (replicated) within-study-comparison, adjusted average total effects can be estimated very similarly across various methods and reasonably close to their randomized (unadjusted) counterpart, although they rest on a functional form assumption (discussed in subsection 1.1.8) as well as on the assumption of Z-conditional unbiasedness (summarized in subsection 1.1.5).

For the non-randomized experiment, a comparison of the pre-treatment scores for students of group C and D reveals that the students in the language group differ from the students in the math group with respect to their abilities in the two measured content domains before the treatment. These observed differences in the pre-test scores, and differences with respect to additional covariates are a common situation in quasi-experimental designs, caused either by (self-) selection or by random fluctuation due to small sample sizes (see Rubin, 1997).[2] For quasi-experimental designs, the effect of these systematic differences would

[2] Adjustment procedures are also suggested to reduce the influence of random fluctuations (see, e. g., Rubin, 2008a, for a recent discussion of this issue).

not vanish, even if we repeated the study, or integrated the results of multiple studies, for example, by meta-analytic techniques (see Campbell & Stanley, 1963, and T. D. Cook & Campbell, 1979, for an early description of the so-called *threats to validity*, to which the biases resulting from non-comparable groups belong to).

From a substantive point of view, self-selection is a very likely reason for the differences in baseline variables between groups in the aforementioned example, as students chose between one of two treatment conditions. There might exist a variety of possible and sound explanations for the observed differences in the pre-test scores. For example, students liking math might prefer the mathematical training as they expect to perform well. Similarly, one could argue that students will select themselves into the mathematical intervention group because they perceive themselves as needy and therefore expect some positive training effects. In any case, from a statistical perspective we cannot infer about the underlying processes that take place "within" the students without further substantive knowledge about the assignment process. In terms of the theory of stochastic causality summarized in the previous chapter, the simple treatment regression $E(Y|X)$ is biased. Although we have not computed any value of an *ATE*–estimator for the treatment effect, we know that the prima facie effect (*PFE*) can not be interpreted as an estimator of the average total effect (*ATE*).

Note that this illustration of a biased prima facie effect (i. e., $PFE \neq ATE$) is not due to the special features of the example presented, where the comparison of two different treatments is analyzed (rather than considering the more natural contrast between treatment and no treatment, see Holland, 1986). But in fact, the same problem would have occurred, had we offered students the choice between participating in the training program and not participating. Furthermore, as Shadish et al. (2008a) argue, the selection mechanism would change, because under these (different) circumstances, the selection of either one of the treatments or the no-treatment condition would be additionally confounded by the different effort required to participate in the study.

The example illustrates that for observational studies it is not guaranteed that treatment groups, or treatment and control groups, are identical with respect to any (pre-treatment) characteristic. If the assignment of units to treatment conditions is beyond the researcher's control, ruling out alternative explanations can be challenging, controversial or even impossible. Without randomization, the average total effect cannot be identified as the simple mean difference of the outcome variable between the treatment groups (in terms of the theory of stochastic causality, the unconditional prima facie effect is not equal to the average total effect). In order to correct the biased prima facie effect, adjustment procedures based on the covariate adjusted means or the Z-conditional propensities rest on the assumption that the covariate-treatment regression for a given vector of covariates Z is unbiased (see section 1.1.5 for details). To evaluate effects of the two training programs, the authors collected a variety of covariates, including pretest measures, demo-

graphics, personality measures, educational background and others (*"The [...] study was designed to have a rich set of covariates potentially related to treatment choice and outcome"*, Shadish et al., 2008a, p. 1340). As mentioned above (see also section 1.1.7), these covariates should be related to students' decisions about whether to select the math or the language course (assignment model) and also to the expected treatment effect of the intervention (outcome model) by substantive theory.

2.2 Review of Adjustment Methods

The literature contains various adjustment procedures for the estimation of average total effects, whose theoretical foundation was summarized in subsection 1.1.8. In order to relate generalized analysis of covariance studied in this thesis to the different research traditions, a short review of different adjustment techniques is given in this section.

We have already described the application of fully saturated modeling of the covariate-treatment regression. If all covariates are discrete (and provided that the covariate-treatment regression is unbiased), the adjusted means as well as their difference can be estimated without further assumptions from the conditional prima facie effects. This approach is known as stratifying on observed covariates (discussed, e. g., by J. Robins & Greenland, 1986; Morgan & Harding, 2006). A similar approach without any functional form assumptions is to *match* units on the observed covariates Z, as described for instance by Cochran (1953) [see, e. g., Rossi, Freeman, & Lipsey, 1999, ch. 13 for empirical applications]. In practice, this nonparametric adjustment method is impossible in finite samples if the dimensionality of the covariates in Z is large, and therefore exact matching pairs cannot be found in the sample (known as *data sparseness problem*, see, e. g., Morgan & Winship, 2007).

To circumvent this problem, alternatives are described in the literature based on distance measures between units with different (multivariate) values of the covariates, for instance, matching based on the Mahalanobis distance (Rubin, 1980). In addition, matching procedures differ with respect to the algorithmic process of matching, whether the matching is performed with or without replacement, the number of matched control units for each treated unit, as well as with respect to other technical features (see, e. g., Gu & Rosenbaum, 1993, for a technical presentation of different matching procedures).

The nonparametric matching on observed covariates is substantially extended through the inclusion of a balancing score based on the assignment modeling approach by Rosenbaum and Rubin (1985). We will discuss the resulting propensity score matching and related methods in section 2.3 (see, e. g., Sekhon, 2009, for a recent survey of matching methods as well as Rubin, 2006, for a collection of relevant work). Note that although matching was originally developed as a nonparametric method without assumptions about the

functional form of a regression, this is not generally true for propensity score based adjustment methods when the Z-conditional propensities have to be estimated (see section 1.1.8 above).

2.2.1 Traditional Analysis of Covariance

An alternative approach to circumvent the data sparseness problem is the assumption of a function form for the covariate-treatment regression. In the most common version (labeled as *traditional analysis of covariance, ANCOVA*) the regression of the outcome variable Y on the covariates Z and the treatment variable X is assumed to be linear (e. g., Cochran, 1957; Rao, 1973; T. D. Cook & Campbell, 1979; Maxwell, Delaney, & O'Callaghan, 1993; Cohen, Cohen, West, & Aiken, 2003, as well as Rubin, 2006, ch. 4):

$$\begin{aligned} E(Y|X,Z) &= \gamma_{00} + \gamma_{01} \cdot Z + \gamma_{10} \cdot X \\ \varepsilon &\equiv Y - E(Y|X,Z). \end{aligned} \tag{2.1}$$

According to the decomposition introduced in subsection 1.1.8 [see Equation (1.26)] the average total effect can be obtained as regression coefficient, i. e., $ATE = E\left(g_1(Z)\right) = E\left(\gamma_{10}\right) = \gamma_{10}$. Typically, the covariate-treatment regression [Equation (2.1)] is estimated by ordinary least-squares. Hence, the traditional ANCOVA is sometimes also simply called *OLS regression* when applied to the estimation of treatment effects (see, e. g., in econometrics Verbeek, 2004, and also Guo & Fraser, 2010).

Statistical Inference The hypothesis of no average total effect in the ANCOVA model is $H_0 : \gamma_{10} = 0$. Additional assumptions are necessary to test this hypothesis for ordinary least-squares estimated covariate-treatment regressions. If these assumptions (particularly with respect to the distribution of the residuals ε, see section 3.2.2) are met, for instance, an F–test based on the general linear hypothesis can be applied (Steyer, 2003, see also subsection 3.2.3).

Example Using this method for the quasi-experimental example means to identify the treatment effect for each treatment (compared to the other) as the regression coefficient γ_{10} [according to Equation (1.26) the effect function $g_1(Z)$] with the pre-treatment covariates as the multivariate covariate $Z \equiv (Z_1, ..., Z_K)$ [included with γ_{10} and γ_{0k} in the intercept function], i. e., for K multiple covariates the traditional ANCOVA–model is $E(Y|X,Z) = \gamma_{00} + \sum_{k=1}^{K}\left(\gamma_{0k} \cdot Z_k\right) + \gamma_{10} \cdot X$. If the assumption of unbiasedness of the covariate-treatment regression as well as the linearity and additivity assumption of this parameterization are met, $\hat{\gamma}_{10}$ is an unbiased estimator of the average total effect. For the estimation of the treatment effect of the math training, the math post-test is used as the outcome variable Y and the math pre-test is one of the covariates. In the same way, the language pre-test is used as one of the covariates for the estimation of the average total effect of the language training, whereas the language post-test is used as the outcome variable. Further covariates used in the study reported by Pohl et al. (submitted) for the estimation of both treatment effects

are demographic variables (gender, age, marital status, major area of study), prior academic achievement (high school grades), topic preference (of math or language) and psychological variables (positive and negative affect). Including additional indicator variables for dummy codings of categorical covariates, Pohl et al. (submitted) used an outcome model according to Equation (2.1) with $K = 33$ covariates.

2.2.2 Analysis of Covariance Without Linearity

The traditional analysis of covariance assumes a linear relation of covariate(s) and outcome (see also subsection 1.1.5). Generalizations exist in the literature which relax this assumption of linearity but still assume additivity (i. e., parallel curves). For example, R. J. A. Little, An, Johanns, and Giordani (2000) applied an extended ANCOVA model without the assumption of linearity for the estimation of an adjusted average total effect. They summarized different data analysis techniques under the following regression equation

$$\begin{aligned} E(Y|X,Z) &= \gamma_{00} + g(Z) + \gamma_{10} \cdot X \\ \varepsilon &\equiv Y - E(Y|X,Z), \end{aligned} \tag{2.2}$$

where $g(Z)$ is a *"smooth" nonlinear function* of the covariates Z. R. J. A. Little et al. (2000) implemented their method based on cubic splines with fixed knots, an analysis that can be conducted with standard program packages for multiple regression if the sample size is large enough. The estimation of $g(Z)$ is based on a polynomial regression model, with additional terms like Z^2 and Z^3. These polynomials can be understood as new covariates, and the parameterization of the covariate-treatment regression is linear in these covariates. The specification of Equation (2.2) implies that there is no interaction between covariates and the treatment variable because X does not enter the nonlinear function $g(Z)$. In other words, because of additivity of effects (see, e. g., Hastie & Tibshirani, 1990), the γ_{10} parameter is still interpreted as an estimator of the average total effect provided that the covariate-treatment regression is unbiased and that the functional form assumption is fulfilled for Equation (2.2) with a specified function $g(Z)$.

Statistical Inference As long as the covariate-treatment regression in Equation (2.2) does not include covariate-treatment interactions and if the regression is estimated by ordinary least-squares, the hypothesis of no average total effect can be tested similarly to the traditional analysis of covariance under the assumptions mentioned above.[3] A test statistic for parallelism of the non-parametric regression curves has been developed by Young and Bowman (1995) and was implemented in R by Bowman and Azzalini (2007).

[3] Note that a generalization of the regression model of Equation (2.2) was presented by J. M. Robins, Mark, and Newey (1992) as *semiparametric causal regression model* (i. e., under the additional assumption of strong ignorability), as an application of semiparametric regression modeling as suggested by Robinson (1988). We do not present semiparametric regression modeling here. For a discussion with respect to nonparametric analysis of covariance see Akritas, Arnold, and Du (2000). Nevertheless, note that the regression presented in Equation (2.2) is a special case of the parameterization of generalized analysis of covariance presented in subsection 1.2 (see also Steyer et al., in press).

Example For the empirical example used in this section to illustrate the different adjustment methods, the function $g(Z)$ could be applied as a flexible model for the regression of the math or language post-test on all pretest measures, i. e., on the multivariate covariates $Z \equiv (Z_1, ..., Z_K)$. For this conditioning of the outcome variable Y on the selected confounders Z with a common nonlinear regression, the estimated coefficient $\hat{\gamma}_{10}$ still has the interpretation as an estimator of the adjusted total treatment effect. Unfortunately, nonlinear analysis of covariance was neither applied by Shadish et al. (2008a) nor by Pohl et al. (submitted), probably because of too small sample sizes (relative to the large number of covariates).

2.2.3 Moderated Regression and Mean-Centering

The regression in Equation (2.2) allows nonlinear dependencies, which are assumed to be parallel between treatment and control group. In order to obtain a model without the assumption of either parallel regression lines or parallel regression curves, a model without additivity can be formulated for the covariate-treatment regression. For the simplest case of two linear regressions conditional on $X = j$, this model is algebraically equivalent to a multiple regression model with interaction terms, also known as moderated regression (e. g., Cohen & Cohen, 1983). From Equation (2.1) we obtain the parameterization for one covariate and a linear relation within each treatment condition by adding the product term $Z \cdot X$ as an additional regressor:

$$\begin{aligned} E(Y|X,Z) &= \gamma_{00} + \gamma_{01} \cdot Z + \gamma_{10} \cdot X + \gamma_{11} \cdot Z \cdot X \\ \varepsilon &\equiv Y - E(Y|X,Z). \end{aligned} \tag{2.3}$$

Due to the interaction, the average total effect no longer equals a single regression coefficient. This follows immediately from Equation (2.3), which fits into the decomposition presented in Equation (1.26) with $g_0(Z) = \gamma_{00} + \gamma_{01} \cdot Z$ and $g_1(Z) = \gamma_{10} + \gamma_{11} \cdot Z$. Hence, the average total effect for a simple model with two treatment groups and one covariate is

$$\begin{aligned} ATE_{10} &= E\big(g_1(Z)\big) \\ &= E(\gamma_{10} + \gamma_{11} \cdot Z) \\ &= \gamma_{10} + \gamma_{11} E(Z). \end{aligned} \tag{2.4}$$

For moderated regression models, an often suggested procedure is to "mean-center" the covariates (see, e. g., Aiken & West, 1996). The appealing improvement of mean-centered covariates is the simple interpretation of γ_{10} as average total effect, if mean-centering yields covariates Z^* with an unconditional expectation of zero, i. e., with $E(Z^*) = 0$ (see also Judd, Kenny, & McClelland, 2001, for the suggestion to center covariates in within-subject designs, Angrist & Pischke, 2009, for a similar suggestion regarding the analysis

of data from a regression discontinuity design, as well as Wooldridge, 2001, for an extended formulation of this idea including the centering for functions of the covariates):

$$\begin{aligned} ATE_{10} &= \gamma_{10} + \gamma_{11} \cdot 0 \\ &= \gamma_{10}. \end{aligned} \tag{2.5}$$

Furthermore, a similar simplification can be obtained for the average treatment effect of the treated, if conditional mean-centering yields $E(Z^*|X=1)=0$.

Statistical Inference Unconditional inference about the average total effect estimated from covariate-treatment regressions with interaction terms (moderated regression models) is discussed in detail in section 3.2.5. To weaken assumptions of ordinary least-squares regressions, we will introduce different implementations within the framework of structural equation modeling. With respect to the mean-centering approach, note that Equation (2.4) and Equation (2.5) deal with the true population value of the covariates' mean.

Example To apply the mean-centering approach for an analysis concerning the given example, a moderated regression with mean-centered covariates $Z^* \equiv (Z_1^*, ..., Z_K^*)$ can be specified by the appropriate linear transformations of each pre-treatment covariate Z_k. With an increasing number of covariates the model becomes more complex, as an interaction term is also included for each additional covariate.[4] For mean-centered covariates, the regression coefficient $\hat{\gamma}_{10}$ is an unbiased estimator of the average total effect, provided that the covariate-treatment regression is Z–conditionally unbiased and that the functional form of $E(Y|X,Z)$ is specified correctly. For moderated regression models without mean-centered covariates, the average total effect is estimated as *average distance* (see section 3.2.4).

Neither Shadish et al. (2008a) nor Pohl et al. (submitted) applied a specification of $E(Y|X,Z)$ with included covariate-treatment interactions. Nevertheless, empirical application of the mean-centering approach for the estimation of causal effects can be found, e. g., in Brand and Halaby (2006) as well as in Zanutto (2006).

2.2.4 Prediction / Regression Estimates

The average total effect can be identified without mean-centering as averages of the difference between regression predictions, a procedure recently suggested by Schafer and Kang (2008) as an alternative to the analysis of covariance. Regression predictions, i. e., *regression estimates* are well known in the survey litera-

[4] Note that even if all covariate-treatment interactions are included, we still rest on assumptions, for example, that no higher order interactions terms, e. g., interactions between covariates are necessary to capture the functional form of $E(Y|X,Z)$ correctly.

ture (see, e. g., Cochran, 1977; Lohr, 1999). The predicted scores incorporated for this approach are obtained from J separate group-specific covariate-regressions

$$\begin{aligned} E_{X=j}(Y|Z) &= \beta_{0j} + \beta_{1j} Z \\ \varepsilon &\equiv Y - E_{X=j}(Y|Z), \end{aligned} \tag{2.6}$$

as $y_{ij} = \beta_{0j} + \beta_{1j} z_i$, using the case-specific value z_i of the covariate Z. For each treatment condition j a predicted score y_{ij} is assigned to each case i, i. e., two predicted scores for each unit under $X = 1$ and $X = 0$ regardless of the observed treatment assignment for the comparison of $J = 2$ treatment groups. Finally, the average total effect is computed as the mean of the differences between the two predicted scores:

$$ATE_{10} = \frac{1}{N} \sum_{i}^{N} \big(y_{i1} - y_{i0}\big). \tag{2.7}$$

The sum in Equation (2.7) is taken over all individuals $i = 1, ..., N$ in the sample and consequently, the observed outcomes under treatment and the observed outcomes under control are replaced by the predicted scores as well.[5] By estimating separate regression models for the treatment group and for the control group, all interactions between the covariates and the treatment variable are included by default. This follows from the fact that the regression coefficients for the J covariate regressions are not constrained to be equal.

Regression estimates, i. e., the estimation of the average total effect as the difference between predicted scores, applies very generally to different kinds of regression models and can therefore be extended very flexibly to model nonlinearities in the covariate-treatment regression. A similar suggestion was made by Wooldridge (2001, p. 609), who pointed out that the conditional regressions $r_j(Z) \equiv E(Y|Z, X = j)$ for each $X = j$ are non-parametrically identified, i. e., these conditional expectations depend entirely on "observables". Hence, when $r_0(Z)$ for treatment $X = 0$ and $r_1(Z)$ for treatment $X = 1$ are known, the *ATE* is identified as

$$ATE_{10} = \frac{1}{N} \sum_{i}^{N} \big(r_1(Z = z_i) - r_0(Z = z_i)\big), \tag{2.9}$$

(see also Imbens, 2004).

[5] Note that for most implementations of the prediction approach, *"the average of the predicted treated outcome for the treated"*, i. e., $\sum_{i=1}^{N} (x_i y_{i1})$ *"is equal to the average observed outcome for the treated"*, i. e., $\sum_{i=1}^{N} (x_i y_i)$ [see Imbens, 2004, p. 12]. Accordingly, for the simple models with linear parameterized intercept and effect functions considered in this thesis, the average total effect in Equation (2.7) is equivalent to

$$ATE_{10} = \frac{1}{N} \sum_{i=1}^{N} \big(x_i(y_i - y_{i0}) - (1 - x_i)(y_i - y_{i1})\big), \tag{2.8}$$

i. e., this approach is equivalent to simple mean imputation (see Schafer & Kang, 2008).

Recently, Glynn and Quinn (2010) applied *generalized additive models* (GAM) [e. g., Hastie & Tibshirani, 1990] for the estimation of average total effects based on predictions of the (group-specific) covariate regression (see the accompanying R-package for details, Glynn & Quinn, 2009).

Statistical Inference Although the resulting estimator of the average total effect for linear parameterized covariate-treatment regressions is equivalent to the estimator based on moderated regressions, the regression estimates discussed in Schafer and Kang (2008) are interesting with respect to the statistical inference. The authors provide adjusted, robust standard errors, which do not assume that the variance of the outcome is modeled correctly (see subsection 3.2.2.1 on page 58).

Note that even though Wooldridge (2001) suggested flexible nonparametric techniques for the estimation of the conditional regressions $E(Y|Z, X = j)$ [for example kernel estimators, Härdle & Linton, 1994], in order to derive valid standard errors, he recommended flexible parametric models based on low-order polynomials (see, e. g., R. J. A. Little et al., 2000, and subsection 2.2.2). Glynn and Quinn (2009) provide bootstrap standard errors for the average total effect.

Example For an application of Schafer and Kang's approach to the quasi-experimental example, the group-specific regressions [Equation (2.6)] could be extended for multiple covariates $Z \equiv (Z_1, ..., Z_K)$ to $E_{X=j}(Y|Z) = \beta_{0j} + \sum_{k=1}^{K} \left(\beta_{kj} Z_k\right)$, for example, under a linearity assumption conditional on $X = j$. Afterwards, a predicted score would be computed for each student under the math treatment and under the language treatment as the weighted sum of the observed values of the covariates z_{ik} and the regression coefficients $\hat{\beta}_{kj}$, estimated by ordinary least-squares:

$$\widehat{y}_{ij} = \hat{\beta}_{0j} + \sum_{k=1}^{K} \left(\hat{\beta}_{kj} z_{ik}\right). \tag{2.10}$$

Finally, the average of the difference between the predicted scores, i. e., $\frac{1}{N} \sum_{i}^{N} \left(\widehat{y}_{i1} - \widehat{y}_{i0}\right)$ is computed as an estimator of the average total effect, $\widehat{ATE}_{10}$.

2.3 Methods based on Propensity Score

The adjustment methods described so far are based on the regression of the outcome variable Y on the observed covariates Z and the treatment variable X. These methods are criticized mainly for the following two reasons: The parametrization of the outcome model (i. e., the functional form of the covariate-treatment regression) is not fixed when the outcome variable Y is analyzed for the first time (Rubin, 2001, 2008b). Therefore, there is a risk that the final result of the analysis (e. g., the estimated average total effect) is taken into account for constructing and specifying the regression model of the (parametric) covariate-treatment regression. Typical post hoc modifications are the adaptation of the functional form of the regression, but

these modifications might also comprise the inclusion or exclusion of covariates, or functions of covariates (for instance, the product of two covariates). Furthermore, although this is of course not specific to the estimation of causal effects, it is often noted, that *typical model-based analyses provide no warning that comparisons may be based on extreme extrapolations* (Rubin, 1997, p. 762, see also, e. g., King & Zeng, 2006; Stürmer, Joshia, Glynn, Avorna, & Rothman, 2006; Zanutto, 2006). We will summarize and discuss these practical issues regarding the adjustment procedures in section 2.5 subsequent to the following presentation of alternative adjustment methods.

Over the last two decades, a set of alternative methods have been developed and become widely accepted and extensively used. These methods are based on the conditional probability of assignment to treatment given the covariates, popularized as *propensity score* by Rosenbaum and Rubin (1983b) [see also Rosenbaum, 2002b, 2002c].[6] As described in section 1.1.8, the theoretical foundation for the identification of the average total effect based on the propensity π is the assumption of Z-conditonal independence of X and the true outcomes, which implies unbiasedness of the regression $E(Y|X,\pi)$.

In quasi-experiments the Z-conditional propensity functions are unknown and have to be estimated. Therefore, adjustment methods based on the propensity score are sometimes called two-step approaches (Dehejia & Wahba, 1999). The estimation of the propensity score for a multivariate vector $Z \equiv (Z_1, ..., Z_K)$ can be performed, for example, by using a parametric logistic regression model (see Rosenbaum & Rubin, 1984),

$$P(X = j|Z) = \frac{\exp\left(\alpha_0 + \sum_{k=1}^{K} \alpha_k \cdot Z_k\right)}{1 + \exp\left(\alpha_0 + \sum_{k=1}^{K} \alpha_k \cdot Z_k\right)}. \tag{2.11}$$

The parameterization of Equation (2.11) represents a functional form assumption, used for the assignment model (as introduced in section 1.1.8). The estimated propensity score $\widehat{\pi}_i$ for unit i is obtained as the predicted score based on the estimated regression coefficients $\widehat{\alpha}_k$ and the observed values z_{ki} for each covariate Z_k. The regression parameters α_k for a logistic regression model are usually estimated by maximum likelihood (see, e. g., Agresti, 2007). It is known that if this assignment model is not specified correctly the average total effect, estimated by conditioning on the estimated propensity score, can be biased (Drake, 1993). That means for the simple logistic model in Equation (2.11), the estimate of the average total effect might be biased if linearity and additivity of the logit $\ln\left[P(X = j|Z)/\left(1 - P(X = j|Z)\right)\right] = \alpha_0 + \sum_{k=1}^{K}(\alpha_k \cdot Z_k)$ do not hold.

There are several alternative strategies for the estimation of propensity scores beyond the simple logistic regression (see, e. g., King & Zeng, 2001, for neuronal networks, McCaffrey, Ridgeway, & Morral, 2004, for a propensity score estimation with boosted regressions, B. K. Lee, Lessler, & Stuart, 2009, for the appli-

[6] Note that in line with Steyer et al. (in press) we distinguish the conditional probability $P(X = j|Z = z)$ [the propensity score for $X = j$ and $Z = z$] and the function $\pi_{X=j} \equiv P(X = j|Z)$ [the *Z-conditional propensity for X = j*] from the true treatment probability functions $P(X = j|Cx)$ [the *true propensity functions*, see section 1.1.6].

cation of classification trees, and Keele, 2008; Woo, Reiter, & Karr, 2008, for a smoothing splines generalized additive model, to name at least a few of them).

For a large number of covariates, the estimation of the regression of the outcome variable Y on all covariates $Z = (Z_1, ..., Z_K)$, the treatment variable X, as well as additional interaction terms (e. g., $Z_1 \cdot X, ..., Z_K \cdot X$) might become complicated due to the problem commonly known as *curse of dimensionality* (see, e. g., Kotz, Read, Balakrishnan, & Vidakovic, 2005, and Hirano & Imbens, 2001). This is especially the case with small samples (for an application of propensity score based adjustment methods for small samples see, e. g., Cepeda, Boston, Farrar, & Strom, 2003). Even if the estimation can be performed without difficulties, redundant regressors might increase the estimated standard errors unnecessarily (because of multicollinearity). Hence, under certain conditions the two-step adjustment based on $\pi_j = P(X = j|Z)$ and $E(Y|X, \pi_j)$ might be easier to model and might give more reliable results than the outcome modeling approach based on the covariate-treatment regression $E(Y|X, Z)$.

The estimated propensity scores can be used in a variety of different data analysis techniques (see, e. g., Guo & Fraser, 2010, for a textbook presentation of different propensity score techniques). Basically, all methods make use of the property that if the treatment assignment is strongly ignorable given the covariates, then the observed difference in the means of the outcome variable between treatment and control group conditional on the propensity score is an unbiased estimator of the conditional total effect given that particular value of the propensity score (see subsection 1.1.8). Note that if the expectation of the conditional total effect given a value of the (estimated) propensity score is taken over the (distribution of the propensity score in the) population of treated units, an estimator of the *average treatment effect of the treated* is constructed (see subsection 1.1.3), whereas the average total effect is obtained with respect to the (distribution of the propensity score in the) total population. The various methods summarized in the following paragraphs differ mainly technically in the way the continuous and uni-dimensional propensity scores are treated for the computation or approximation of the expectation.

Example In the examples of Shadish et al. (2008a) and Pohl et al. (submitted), the propensity scores were estimated by using logistic regressions. The procedures were designed and applied in order to achieve acceptable balance of the considered covariates $Z \equiv (Z_1, ..., Z_K)$, a property we will discuss in subsection 2.5. Furthermore, Luellen, Shadish, and Clark (2005) compared three approaches for the construction of propensity scores for data from the same within-study comparison: classification tree analysis, bagging classification trees and simple logistic regression.

2.3.1 Propensity Score Subclassification

A very fundamental approach for the adjustment based on the estimated propensity scores is the creation of S strata, in which the units of the treatment group and the control group have identical (or at least "comparable") propensity scores. In order to apply the *propensity score subclassification* approach[7] as an adjustment method, a new variable π_s is created, for instance with approximately equal sized classes based on the percentiles of the estimated propensity score distribution. For this purpose, a corresponding value $1,...,S$ — according to the estimated propensity score percentile — is assigned to the variable π_s for each unit $U = u$, regardless of the observed value of the treatment variable X. The underlying idea can be described as conditional mean adjustment, where the adjusted means $E\left(E_{X=j}(Y|\pi)\right)$ are identified (or at least approximated) by the expectation of $E_{X=j}(Y|\pi_s = s)$, for each of the $S = s$ strata and each treatment group $X = j$, over the marginal distribution of π_s. The difference between the two adjusted means for group $X = j$ and $X = k$ yields an estimator of the average total effect ATE_{jk} (see, Steyer et al., in press).

Rosenbaum and Rubin (1984) suggest forming $S = 5$ strata, because Cochran (1968a) reported a 90 % bias reduction for the adjustment by subclassification for a single confounding covariate with 5 strata. This finding was replicated for propensity score subclassification by Drake (1993), who also pointed out that *"there was some bias left in estimating treatment effect when using the propensity score approach in the continuous model"*. To remove this (remaining) bias, which is sometimes called *residual confounding* (e. g., Austin & Mamdani, 2006), the combination of different adjustment procedures is discussed in the literature (see subsection 2.4).

For the estimation of the conditional expectations $E_{X=j}(Y|\pi_S = s)$ it is necessary that at least one unit be observed with an estimated propensity score in each stratum s for each treatment group j. In other words, the number of strata has to be small enough that the conditional mean differences $\widehat{PFE}_{10;\pi_s=s}$ — used as estimators of $CTE_{10;\pi_s=s}$ — are empirically identified. Hence, the optimal number of strata depends on the data. For a small dataset it may not be possible to use five strata (Rubin, 1997), whereas for a large dataset more than five propensity score subclasses may be desirable (Huppler Hullsiek & Louis, 2002). For π_s as introduced above, the conditional total effect of treatment $X = j$ compared to treatment $X = k$ for each stratum is expressed as:

$$CTE_{jk;\pi_s=s} = E_{X=j}(Y|\pi_s = s) - E_{X=k}(Y|\pi_s = s). \tag{2.12}$$

[7] Note that this approach is described under different names, i. e., beside *propensity score subclassification*, for instance, *propensity score stratification* (e. g., Senn, Graf, & Caputo, 2007), and *blocking on the propensity score* (e. g., Imbens, 2004) are commonly used.

The average total effect is identified as the weighted sum over the strata-specific conditional total effects,

$$ATE_{jk} = \sum_{s=1}^{S} \left(\frac{N_s}{N} CTE_{jk;\pi_s=s} \right), \tag{2.13}$$

with N_s as the number of units within each stratum and $N = \sum_{s=1}^{S}(N_s)$ as the total sample size. From $\frac{N_s}{N}$ it follows that the weighting is proportional to the number of observations falling in each stratum.

Statistical Inference Standard errors for the propensity score subclassification *ATE*–estimator are commonly calculated as

$$Var\left(\widehat{ATE}_{10}\right) = \sum_{s=1}^{S} \left[\left(\frac{N_s}{N} \right)^2 Var\left(\widehat{CTE}_{10;\pi_s=s}\right) \right], \tag{2.14}$$

i. e., based on the variance for the conditional total effect estimator given $\pi_s = s$ (see, e. g., Benjamin, 2003, and Guo & Fraser, 2010, p. 66). As discussed by Zanutto (2006), these standard errors are only approximations, because they do not account for the fact that subclassification is based on estimated propensity scores. Alternative standard errors exist, for example, based on bootstrapping confidence intervals (see, e. g., Tu & Zhou, 2003).

Example For the quasi-experimental study used as an illustration of different adjustment methods, Shadish et al. (2008a) applied the subclassification approach on the estimated propensity score as described in this paragraph and reported standard errors from $N = 1000$ bootstrap samples. Due to a smaller sample size, Pohl et al. (submitted) did not perform the propensity score subclassification as described here. Instead, a weighted regression analysis was performed with individual weights obtained from the stratum membership for each unit. This is a special version of *inverse-propensity weighting* with coarsened weights as described in the next paragraph (see, e. g., Freedman & Berk, 2008, for the similarity of simple weighted regression and inverse-propensity weighting).

2.3.2 Inverse-Propensity Weighting

Propensity score subclassification can be understood as a special case of a more general weighting approach (Lunceford & Davidian, 2004),[8] and the relation of both approaches can be described as *bias-variance trade-off* (Tan, 2007): On the one hand the bias of the (approximated) *ATE*–estimator decreases with an increasing number of strata, but on the other hand the variance of the *ATE*–estimator (i. e., the standard error) decreases with an increasing sample size per stratum. A similar approach to inverse-propensity weighting is known in survey sampling (see, e. g., Kish, 1965), and the estimator of the average total effect can be ex-

[8] Also the reverse was argued, that *"... weighting can also be viewed as the limit of subclassification as the number of observations and subclasses tend to infinity."* (see Rubin, 2001, p. 173).

pressed as the difference of two Horvitz-Thompson estimators of the adjusted means (Horvitz & Thompson, 1952). The average total effect based on the weighted outcome is defined as

$$\begin{aligned} ATE_{10} &= E(\tau_1) - E(\tau_0) \\ &= E_{X=1}\big[E_{X=1}(Y_W|Z)\big] - E_{X=0}\big[E_{X=0}(Y_W|Z)\big] \\ &= E(Y_W|X=1) - E(Y_W|X=0), \end{aligned} \tag{2.15}$$

with $Y_W \equiv Y \cdot \left(X \cdot \frac{P(X=1)}{P(X=1|Z)} + (1-X) \cdot \frac{1-P(X=1)}{1-P(X=1|Z)}\right)$ [see Steyer et al., in press]. Estimates of the average total effect based on weighting are obtained by replacing the true Z-conditional probabilities $P(X = j|Z)$ with the estimated propensity scores $\hat{\pi}$. The propensity scores are estimated under the assumption that the assignment model is correctly specified (see above).[9] Different `R` implementations for various weighting estimators are available (see, e. g., Ridgeway, McCaffrey, & Morral, 2006; Glynn & Quinn, 2009).

It is well known in the literature that inverse-propensity weighting is susceptible for extreme propensity scores: *"This process [inverse-propensity weighting] can generate unrealistically extreme weights when an estimated propensity score is near zero or one, something that is avoided in the subclassification approach"* (Rubin, 2001, p. 184). This makes inverse-propensity weighting prone to misspecified propensity score models (Kang & Schafer, 2007a). Propensity score subclassification does not incorporate the raw estimated treatment probabilities for the computation of the conditional average total effect $CTE_{10;\pi_s=s}$ and is therefore, compared to the weighting approach, more robust against misspecifications of the assignment model (see also R. J. A. Little & Rubin, 2002).

Statistical Inference In order to obtain a valid standard error for the estimated average total effect based on inverse-propensity weighting, it is necessary to account for the fact that the weights are based on the propensity scores $\hat{\pi}$ estimated from the sample. Unfortunately *"weights are routinely treated as known by nearly all researchers who use them"* (Morgan & Todd, 2008, p. 278). Nevertheless, correct standard errors for a specific *ATE*–estimator obtained from inverse-propensity weighting with propensity scores estimated by a logistic regression [see Equation (2.11)] were recently derived by Schafer and Kang (2008).

[9] Note that two versions of the inverse-propensity weighting estimator exist in the literature which differ with respect to the normalization of the weights

$$\widehat{ATE}_{10} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i \cdot y_i \cdot \frac{1}{\hat{\pi}_i}\right) - \frac{1}{N}\sum_{i=1}^{N}\left((1-x_i) \cdot y_i \cdot \frac{1}{1-\hat{\pi}_i}\right) \tag{2.16}$$

(see, e. g., Rosenbaum, 1987, 1998) or

$$\widehat{ATE}_{10} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i \cdot y_i}{\hat{\pi}_i}\right) / \left(\sum_{i=1}^{N}(x_i/\hat{\pi}_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(1-x_i)\cdot y_i}{1-\hat{\pi}_i}\right) / \left(\sum_{i=1}^{N}(1-x_i/1-\hat{\pi}_i)\right). \tag{2.17}$$

(see, e. g., Hirano & Imbens, 2001; Tan, 2007). Both procedures are assumed to be equal for large sample sizes, because the two correction terms have an expectations of one (a formal argumentation can be found, e. g., in Lunceford & Davidian, 2004).

Example Shadish et al. (2008a) applied inverse-propensity weighting as well as weighting in combination with adjustment for additional covariates as described by Rubin (2001). In order to obtain standard errors for the *ATE*–estimator, the authors used a bootstrapping approach again.

2.3.3 Propensity Score Matching

Propensity score matching describes a class of methods dealing with pairing units from the treatment group and the control group with equal (or at least "similar") values of the estimated propensity scores. Subclassification and propensity score matching share the same idea of comparing treated and untreated units conditional on the estimated propensity score, i. e., both approaches are based on the assumption of unbiasedness of $E(Y|X,\pi)$. Unlike the propensity subclassification approach, propensity score matching is applied by the selection of a matched sample of treated and untreated units with similar propensity score distributions instead of dividing the sample into S distinct strata with similar values of the propensity scores. Originally, matching was described as *"sampling from a large reservoir of potential controls to produce a control group of modest size in which the distribution of covariates is similar to the distribution in the treated group"* (Rosenbaum & Rubin, 1983b, p. 48). Hence, propensity score matching is asymmetric in the sense that unmatched units (from the larger group) are typically discarded for the estimation of adjusted treatment effects based on propensity score matched samples (see, e. g., Schafer & Kang, 2008, for some comments about the consequences). This is an implicit restriction of the analyzed sample to the region of common support (see, e. g., Guo & Fraser, 2010, and subsection 2.5).

Nonparametric matching on the observed covariates (as described above) and matching on the (estimated) propensity score is often performed based on the same matching algorithms. Different matching techniques can be described again as a trade-off between *incomplete* matching (i. e., failing to match all treated subjects to control units) versus *inexact* matching (i. e., the matching of dissimilar, not comparable subjects).

Matching on propensity scores and matching on observed covariates are often combined, a strategy described as the observational study analog of blocking in a randomized experiment (Rubin & Thomas, 2000). For instance, one of the most common techniques is the one-one matching of units based on the minimal Mahalanobis distance of the observed covariates $Z = (Z_1, ..., Z_K)$, where the matching is restricted to units within a range of the estimated propensity score logit [$\hat{\pi}_{ln} = \ln\left(\hat{\pi}/(1-\hat{\pi})\right)$, see Schafer & Kang, 2008, for a short summary of this matching procedure]. For this matching strategy, c is called the *caliper*, a pre-specified tolerance in terms of the propensity, and observed covariate matching is performed only in the range $\hat{\pi}_{ln} \pm c$ (see Rosenbaum & Rubin, 1985, for the selection of value for the caliper). A survey of all different matching estimators is beyond the scope of this review (see, e. g., Rubin & Thomas, 1996,

D'Agostino, 1998, Rosenbaum, 2002b, Imbens, 2004, and Baser, 2006, for a survey of techniques). Gu and Rosenbaum (1993) compare different multivariate matching methods and Froelich (2004) gives a review of the small sample properties of various matching estimators. The choice between different propensity score matching estimators can be made based on the distribution of the propensity scores in the treatment group and the control group (see Dehejia & Wahba, 2002, for details). Finally, see Ho, Imai, King, and Stuart (2005) for an implementation of the Mahalanobis matching in `R`, Leuven and Sianesi (2003) for a `Stata` package and Parsons (2004) for an implementation of different matching estimators in `SAS`.

Statistical Inference The *ATE*–estimator based on the propensity score matched samples can be computed as a simple mean difference. A simple t-test for unpaired samples is suggested to test the hypothesis of no average total effect (see, e. g., Schafer & Kang, 2008). This procedure is valid only for known propensity scores (Imbens, 2004), a limitation which is often ignored in applied research.[10]

Bootstrapping approaches are implemented for the computation of standard errors for the average total effect based on matching estimators (see, e. g., Becker & Ichino, 2002), but Abadie and Imbens (2006) showed that bootstrapping yields biased variance estimates for matching estimators. As discussed, e. g., by Stuart (2008), with respect to valid standard errors for matching estimators there is no consensus in the statistical literature (see also Hill & Reiter, 2006).

Finally, note that an interesting conclusion was drawn by Ho, Imai, King, and Stuart (2007, p. 224, nomenclature changed), who described matching as nonparametric preprocessing to reduce model dependence (i. e., to make the adjustment more robust against misspecified functional form assumptions for the outcome model or the assignment model): *"We thus take advantage of a common feature of all the methods of computing uncertainty estimates associated with regression-type parametric methods: They are all conditional on the pretreatment variables Z (and X), which are therefore treated as fixed and exogenous. Since our preprocessing procedures modify the raw data only in ways that are solely a function of Z, a reasonable method for defining uncertainty is to continue to treat Z, and thus our entire preprocessing procedures, as fixed".* In section 3.2.5 we will derive under which conditions the mentioned assumption of fixed regressors (i. e., the treatment of covariates and the treatment variable as non-stochastic regressors) yields unbiased standard errors of the average total effect.

[10] See, for instance, Gelman and Hill (2007, p. 212): *"The standard errors presented for the analyses fitted to matched samples are not technically correct. First, matching induces correlation among the matched observations. The regression model, however, if correctly specified, should account for this by including the variables used to match. Second, our uncertainty about the true propensity score is not reflected in our calculations. This issue has no perfect solution to date and is currently under investigation by researchers in this field."*

2.3.4 Propensity Score ANCOVA

As summarized in section 1.1.6, Z-conditional independence of X and the true outcomes implies unbiasedness of the regression $E(Y|X,\pi)$ [see also Rosenbaum and Rubin (1983b), and Steyer et al., in press]. In line with this implication, e. g., Dehejia and Wahba (1999) note that besides propensity score subclassification and propensity score matching, $E(Y|X = j,\pi)$ for each j in X can be estimated with various methods for flexible functional forms of regressions, for instance, by standard non-parametric techniques (e. g., Härdle, 1990; Cantoni & De Luna, 2006).

Example Shadish et al. (2008a) computed an *ATE*–estimator based on parametric models for $E(Y|X,\pi)$ and $E(Y|X,Z,\pi)$, and labeled the procedures as *propensity score ANCOVA*. The estimated propensity score $\hat{\pi}_{ln}$ (logit transformed, see above) was included as an additional linear predictor, and also as a nonlinear (quadratic and cubic) predictor in the regression of the outcome variable Y on the multivariate covariates $Z = (Z_1,...,Z_K)$ and the treatment variable X.

2.4 Combined Methods

Although outcome modeling and assignment modeling appear to be competing approaches, different methodological developments combine both strategies. We will briefly discuss two basic goals in this direction which raise further research questions and possible extensions for the structural equation models with nonlinear constrain which are studied in this thesis (see subsection 5.4.2).

2.4.1 Gaining Efficiency

The combination of an outcome model and an assignment model is suggested to gain efficiency of the propensity score based adjustment methods. For instance, regression adjustment can be applied to reduce *residual within-stratum confounding* subsequently to propensity score subclassification (see, e. g., for empirical applications Kurth et al., 2006). As described by Rubin (2001), a combined method can be constructed by replacing the stratum-specific difference $E_{X=1}(Y|\pi_s = s) - E_{X=0}(Y|\pi_s = s)$ in Equation (2.12) with an additional regression adjustment for the observed covariates Z, i. e.,

$$CTE_{10;\pi_s=s} = E_{X=1}(Y|\pi_s = s,Z) - E_{X=0}(Y|\pi_s = s,Z) \tag{2.18}$$

[see, e. g., Lunceford & Davidian, 2004, for an application where the regression $E_{X=j}(Y|\pi_s = s,Z)$ is specified as linear regression with the same parameters for each stratum $S = s$ and for each treatment group j].

For propensity score matched samples, a similar strategy is known as *postmatching analysis* (see, e. g., Guo & Fraser, 2010). For instance, Rosenbaum (1986) applied an analysis of covariance on matched pair

differences. Alternatively, inverse-propensity weighting and regression adjustment can be combined to gain efficiency of an adjustment procedure (see Hirano & Imbens, 2001).

2.4.2 Double Robustness

To conclude this survey of adjustment methods for quasi-experimental designs, we must finally mention a special class of dual-modeling strategies. These recently suggested estimators for the average total effect were not developed to gain efficiency of the adjustment procedure, but to make the resulting analysis more robust against the misspecification of either the assignment model or the outcome model. These so-called *doubly robust* procedures (see Lunceford & Davidian, 2004; Bang & Robins, 2005; Kang & Schafer, 2007a) are also known in the missing value literature (see, e. g., J. M. Robins & Rotnitzky, 1995; R. J. A. Little & An, 2004; Carpenter, Kenward, & Vansteelandt, 2006). Kang and Schafer (2007a) discuss the performance of different doubly robust strategies for the estimation of the average total effect. Until today, there has been an ongoing discussion about the usefulness of doubly robust estimation methods in the literature (see, for example, the comments to Kang & Schafer, 2007a, by J. M. Robins, Sued, Lei-Gomez, & Rotnitzky, 2007; Ridgeway & McCaffrey, 2007; Tan, 2007; Tsiatis & Davidian, 2007). The variance of the *ATE*–estimator is increased for doubly robust methods compared to the procedures based either solely on the assignment or solely on the outcome model (see, e. g., Freedman & Berk, 2008). Hence, doubly robust estimators are clearly suboptimal, if one of the parameterizations of the function form of the outcome or the assignment model is (precisely) correct. For a detailed description of the different doubly robust *ATE*–estimators and their theoretical foundation see, e. g., Lunceford and Davidian (2004), Tan (2006), Tsiatis (2006), and Glynn and Quinn (2010).

To analyze the data from the described quasi-experimental design, Pohl et al. (submitted) applied a weighted covariate-treatment regression with a selection of covariates and weights according to a propensity score subclassification. This analysis can be seen as double robust, provided the covariate-treatment regression is assumed to be Z–conditionally unbiased with the selection of potential confounders included as covariates Z.

2.5 Practical Issues Concerning the Adjustment

As described above, the average total effect can be estimated unbiasedly if a) the covariate-treatment regression is unbiased and b) either the covariate-treatment regressions (outcome modeling) or the propensity score model (assignment modeling) is precisely correctly parameterized (see subsection 1.1.8). However, in practice some additional properties of the adjustment methods, which we will highlight in this subsection, are relevant.

2.5.1 Balancing Check

A key concept for the practical application of adjustment methods based on the propensity score is the inspection of the achieved *balance* of the covariate-distributions of the matched samples or the weighted covariates. Rosenbaum and Rubin (1983b) have shown that the covariates are independent of the treatment variable conditional on the true propensity, i. e.,

$$Z \perp\!\!\!\perp X|\phi. \tag{2.19}$$

(see also Steyer et al., in press). This balancing property of the true propensity (which relies on large samples for the estimated propensity score) can be used to evaluate the specification of the assignment model used for the estimation of the propensity score $\widehat{\pi}$, provided Z-conditional independence of X and the true outcomes. If Equation (2.19) is not fulfilled for an estimated propensity score, the assignment model for $\pi_j = P(X = j|Z)$ cannot be specified correctly. Obviously, the propensity score $\pi_j = P(X = j|Z)$ is modeled without taking the outcome variable Y of the study into account (see above). Hence, it is possible (and also strongly suggested, see, e. g., Dehejia, 2005) that different specifications of the assignment model be tried as often as necessary to obtain trustable estimates of the propensity score in terms of balance (see also Yanovitzkya, Zanutto, & Hornik, 2005, for a step-by-step example).

According to Equation (2.19), treated and untreated units have identical distributions of the covariates $Z = (Z_1, ..., Z_K)$ for a given value of the true propensity score. Therefore, we expect the distributions of the covariates conditional on the propensity score $\pi = c$ to be equal between the treatment groups $X = j$ and $X = k$ for all values of c between 0 and 1 (see Schafer & Kang, 2008). For the univariate distribution this implies that, for instance, the first two central moments are equal conditional on the values $\pi = c$:

$$\begin{aligned} E(Z|\pi = c, X = j) &= E(Z|\pi = c, X = k) \\ Var\,(Z|\pi = c, X = j) &= Var\,(Z|\pi = c, X = k). \end{aligned} \tag{2.20}$$

In order to analyze the achieved balance simultaneously for many values c of the estimated propensity score $\widehat{\pi}$ under realistic sample sizes, Rosenbaum and Rubin (1984) suggest an approach based on propensity score subclassification (i. e., a two-way analysis of variance with treatment variable X times π_s as an indicator for the propensity subclass, conducted separately for each covariate). Alternatively, the weighted covariates can be analyzed to investigate the balance property, as described, for instance, by Steyer et al. (in press).

The assignment model itself is — from a substantive point of view — not of interest for the estimation of average total effects. This fact has motivated the conclusion that the assignment model need not be

parsimonious and might include numerous covariates, interactions and nonlinear terms (see, e. g., Shaha, Laupacisa, Huxa, & Austina, 2005). Hence, given that the assignment model itself is only used as a device for estimating the (individual) propensity scores, statistical significance of regression coefficients in the logistic regression are usually not of great interest for the model building process (see subsection 2.5); Rosenbaum and Rubin (1984) therefore suggest using a liberal inclusion criterium for the (stepwise) regression. Furthermore, as pointed out by Imai, King, and Stuart (2008), if the balance is checked for matched samples, the sample size of the comparison group (related to the number of controls dropped through the matching process) can interfere with the test statistic used for the balance check.

Although the balance check is an important step for the specification of the assignment model, it is often found to be poorly implemented or even omitted in applied studies using propensity score techniques (see, e. g., Austin, 2008). Nevertheless, basic criteria for applied researchers to check the achieved balance are provided by Rubin (2001). In line with these rules, Pohl et al. (submitted) checked the specification of the propensity score model for the observational study (described as an example in section 2.1) according to the following two main balance metrics: The absolute standardized difference in the covariates' means (Rubin, 1973; Rosenbaum & Rubin, 1985; Haviland, Nagin, & Rosenbaum, 2007) and the variance ratio of the covariates between treatment groups.

2.5.2 Regression Diagnostic

The corresponding counterpart to the balance check for the specification of the assignment model is the application of *regression diagnostics* for an evaluation of the parameterization of the outcome model. If the functional form of the covariate-treatment regression is misspecified, the *ATE*–estimator will be biased even if the covariate-treatment regression is Z-conditional unbiased. Therefore, regression diagnostics is of major importance for the outcome modeling approach.

Various regression diagnostics are known in the literature: For instance, Steyer (2003) describes a test statistic based on a comparison of a parametric regression model (with a functional form assumption) to a saturated model, which is obtained by creating indicator variables for each distinct pattern of values of the regressors. Therefore, this strategy can be applied for sufficiently large samples and for discrete covariates. If no saturated parametrization is possible to test the functional form assumption of the covariate-treatment regression, the specification can be evaluated by comparing different parameterizations of a regression, for example, a simple linear functional form against an extended model including additional power terms (a strategy known as *RESET* test developed by Ramsey, 1969).

According to Long and Trivedi (1992), possible misspecifications can be categorized as first order misspecifactions regarding the conditional expectation of the residual $[E(\varepsilon|X) \neq 0]$, as second order misspec-

ifications regarding the structure of the error variances [$Var(\varepsilon|X) = E(\varepsilon^2|X) \neq \sigma^2 I$] or as third order misspecifications regarding the distribution of the error terms. This classification is of interest for the implementation of generalized analysis of covariance because opposed to the general linear model some of the studied structural equation models with nonlinear constraints are with respect to the *ATE*–estimator not robust against second order misspecifications.

Most commonly, the specification of a regression model is judged based on a (visual) inspection of the residuals, a strategy known as *residual analysis* (see, for example, Belsley, Kuh, & Welsch, 1980, Kutner, Nachtsheim, Neter, & Li, 2005, Kang & Schafer, 2007b, and Sheather, 2009, as well as Tan, 2007, for a discussion of differences in model checking between outcome regression and propensity score specification). Since the residuals of a true regression are uncorrelated with any functions of the regressors and have an expectation of zero conditional on (functions of) the regressors, the inspection of (standardized) residual plots can help to detect inappropriate specifications of the covariate-treatment regression (see, for example, S. Weisberg, 2005).

2.5.3 Overlap, Common Support & Extrapolation

Within the Rubin Causal Model, substantial overlapping distributions of the covariates or the propensity scores between treatment groups is discussed as a desirable property for causal inference from observational data. For example, Rubin (2001, p.176) warns, that *"when there are some treated subjects with propensity scores outside the range of the control subjects, no inference can be drawn concerning the effect of treatment exposure for these treated subjects from the data set without invoking heroic modeling assumptions based on extrapolation."*

Without a parametric model assumption for the regression $E(Y|X,\pi)$ overlapping distributions are technically needed for some of the propensity score approaches (see, for instance, the propensity score subclassification described above). For adjustment methods based on parametric covariate-treatment regressions — like the analysis of covariance and the moderated multiple regression approach — sufficiently overlapping distributions (of covariates or propensity scores) among treated and untreated units are technically unnecessary. If the functional form assumption of the regression $E(Y|X,Z)$ is precisely correct (and if the covariate-treatment regression is Z–conditionally unbiased), and if the individual treatment probabilities are different from zero and different from one for each treatment condition and for each unit [that means for the two group case $0 < P(X = 1|U = u) < 1$ for each unit $U = u$], then the average total effect based on extrapolation is well defined and can be unbiasedly estimated.

Additionally, as Kang and Schafer (2007b, p. 575) summarized, extrapolation is a common phenomenon for the estimation of causal effects from quasi-experimental data: *"All of our methods extrapolate. The as-*

sumption of ignorability is itself an extrapolation." Nevertheless, (residual) diagnosis for the functional form assumption of the covariate-treatment regression can only be performed empirically within the observed range of data (see above). Furthermore, the average total effect is estimated as the expectation over the unconditional distribution of the covariates. It seems therefore reasonable to conclude that the extrapolation might make the estimator of the average total effect unstable and vulnerable to misspecifications of the utilized parametric regression model (see, e. g., King & Zeng, 2006).

It should be noted that this is not necessarily a major drawback. If a missing overlap is observed by checking the empirical distribution of the weighted covariates or the estimated propensity scores, one possible solution might be the estimation of the average total effect of the treated (provided that at least the estimated distribution for the treated group overlaps with the distribution of the control group). Furthermore, similar to the common strategy for propensity score based analysis, the target population for which the average total effects can be estimated from a concrete observational study might be reduced for the outcome modeling approach as well as for the overlapping cases (Morgan & Winship, 2007).

2.5.4 Unmeasured Confounders

The different adjustment methods reviewed in this section are based on the assumption of unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ and are therefore restricted to measured potential confounders. Unmeasured confounders or omitted covariates are also treated as a problem of *selection based on unobservable variables* (see, e. g., Heckman & Robb, 1985). Often, data on important confounders are not available in practical applications. Angrist and Krueger (1999) suggest checking the sensitivity of the estimated treatment effects to changes in the set of included covariates $Z \equiv (Z_1, ..., Z_K)$ as a hint that unobserved covariates would change estimates further.

An elaborated strategy that deals with the violations of the fundamental unbiasedness assumption of the covariate-treatment regression is an approach known as *sensitivity analysis* (see, e. g., Rosenbaum & Rubin, 1983a; Lin, Psaty, & Kronmal, 1998, or Rosenbaum, 2005, for an introduction). Sensitivity analysis addresses the question of how hidden biases due to unmeasured confounders of various magnitudes might alter the final interpretation of the results from an observational study. This is performed with respect to the qualitative conclusions drawn based on the estimated average total effect (see Rosenbaum, 2002b, for an extensive presentation of the various methods and their theoretical foundation).

2.5.5 Measurement Error

We will now briefly discuss the most distinct feature of generalized analysis of covariance implemented as structural equation model with nonlinear constraints: The flexibility to take the measurement error of covariates into account for the estimation of a latent covariate-treatment regression. As recently demon-

strated by T. D. Cook, Steiner, and Pohl (2009), the reliability of the selected covariates (see section 1.1.7) is important to obtain unbiased estimates of the average total effect. Various adjustment methods, for instance, matching on covariates are prone to measurement error (Shadish & Cook, 2009). The same is true for stratifying on observed covariates.

For the outcome modeling it is well known that estimates of regression coefficients in manifest regression models are underestimated if regressors are measured with measurement error (see, for example, Lord, 1960; Cochran, 1968b; Degracie & Fuller, 1972; H. I. Weisberg, 1979; Cochran, 1983; Fuller, 1987; Angrist & Krueger, 1999). Therefore, biased (also known as *attenuated*) regression coefficients might restrict the use of the discussed adjustment methods to purely descriptive purposes when the biased regression coefficients lead to biased *ATE*–estimates. Hence, covariates measured with error (also known as *fallible covariates*) can be seen as a major drawback, as summarized by Bentler (1991, p. 159): *"The standard method for doing this, analysis of covariance (ANCOVA), may fail to give an unbiased treatment effect because if the control variables are fallible, the control is at the level of an observed variable rather than at the level of the true characteristic"* (see also the discussion in T. D. Cook & Campbell, 1979, ch. 4, and Huitema, 1980, p. 149–154, as well as, e. g., Bini, Monari, Piccolo, & Salmaso, 2010, for measurement error and value added modeling). Similar effects of measurement error are also well known for the analysis of change without explicitly referring to the estimation of causal effects (see, e. g., Cribbie & Jamieson, 2000; Jamieson, 2004; Cribbie & Jamieson, 2004).

Different adjustment methods have been suggested for dealing with measurement error. For instance, Sörbom (1978) suggested an alternative to the analysis of covariance without interactions based on the group-specific regression (see also Arbuckle, 2006, for a technical description of this method as well as Aiken, Stein, & Bentler, 1994, for an application). Furthermore, West, Biesanz, and Pitts (2000) describe an analysis of covariance that corrects for unreliability of the covariates, without the specification of a measurement model.

Generalized analysis of covariance for the estimation of average total effects implemented as structural equation model with nonlinear constraints can be applied straight forward if covariates are measured with error and an appropriate measurement model can be formulated and identified (see Steyer & Partchev, 2008, and Steyer et al., in press). It will not be necessary to consider measurement models for latent covariates with respect to the research questions focused on this thesis. Nevertheless, we will give a summary of subsequent research questions dealing with measurement error of covariates in section 5.4.2.

2.5.6 Robustness against Misspecifications of the Functional Form Assumption

Given the two more general alternatives of outcome modeling versus assignment modeling, the robustness of the adjustment procedures against possible misspecifications of the functional form of either the regression Y on X and Z or the regression X on Z is probably the most interesting question for applied researchers. Even though a Monte Carlo simulation by Drake (1993) is often cited for a comparison of both approaches with respect to robustness against misspecifications, the generalization of Drake's results is questionable, because for empirical applications neither the true model for the outcome nor the true assignment model are known. Hence, the consequences of model misspecification are subtle and elusive, as well as difficult to investigate in a simulation study.

Besides theoretical considerations regarding the adjustment methods, empirical comparisons are a valuable way for illustrating how well certain adjustment procedures for data from quasi-experimental designs yield unbiased estimated average total effects. Shaha et al. (2005) concluded from a systematic review of published observational studies which used outcome modeling as well as assignment modeling *"that the two methods usually did not differ in the strength or statistical significance of associations between exposures and outcomes"*. Alternatively, the integration of different studies for the evaluation of a specific kind of treatment in a given content domain (based on randomized experiments and quasi-experimental designs) could be performed, for instance, with the help of meta-analytic techniques. Instead of reviewing the results of those meta-analyses, we will present the results of the more elegant *within-study comparisons* in the subsequent subsection. As mentioned above, within-study comparisons take the observed average total effect from a randomized experiment and contrast it with the adjusted average total effect from an observational study which shares *the same treatment* (T. D. Cook et al., 2008).

2.6 Performance of the Adjustment Methods in the Example

In section 2.1 we described a simple quasi-experimental design to evaluate two training programs. Most of the adjustment strategies reviewed in the last subsection were applied to a study with the introduced quasi-experimental design, published by Luellen et al. (2005), Shadish et al. (2008a), and replicated by Pohl et al. (submitted). In order to judge the appropriateness of the estimated (adjusted) causal effects from the non-randomized study, the authors actually used an extended design, where a direct comparison between randomized and non-randomized conditions is possible. Figure 2.2 repeats the complete design of the *within-study comparison*.

In a first step, the $N = 445$ sampled undergraduate students ($N = 202$ in the replication of Pohl et al., submitted) were randomly assigned to one of the following two conditions: A non-randomized experiment

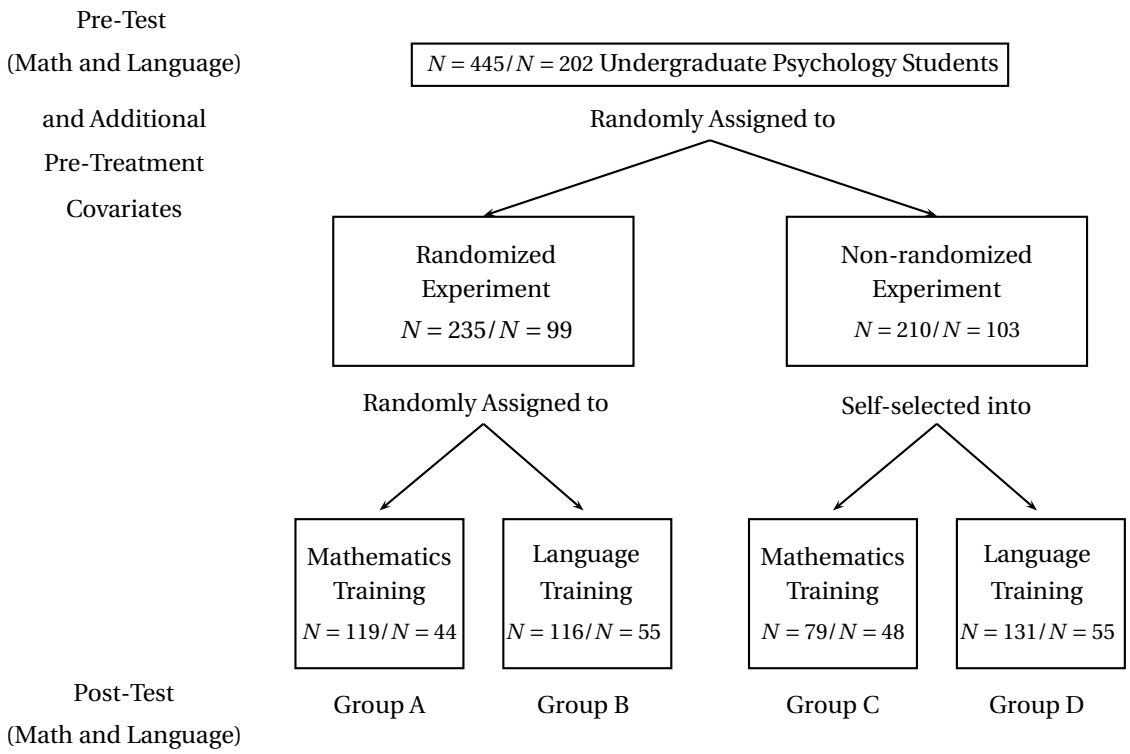


Figure 2.2: Design of the complete within-study comparison for the introductory example

(see section 2.1) and a randomized experiment with exactly the same interventions. The random assignment to one of the two "arms" of the design was conducted after measuring the pre-treatment baseline covariates. Within the randomized condition of the within-study comparison, students were furthermore randomly assigned to the mathematics or the language treatment group. In the non-randomized arm, students chose the training they wished to attend (as described above). Regardless of their assignment condition, all students attended exactly the same training courses either in mathematics (groups A and D) or in English (groups C and D), at the same time and in the same setting. Finally, all students answered the same battery of questions as well as the math and language test as outcome measures. In other words, the non-randomized arm and the randomized arm of the design are absolutely comparable, except for the different assignment mechanisms in the second step of the design (see Figure 2.2).

The ingenious design of the within-study comparison allows the performance of the adjustment methods to be described in a meaningful way. If randomization was applied successfully, we can interpret the results of the randomized arm as unbiased estimates of the average total effects of the training. The authors therefore compute the *percent bias reduction* of the estimated average total effect from the non-experimental arm compared to the credible estimate based on the truly randomized trial.

A comprehensive presentation of the results for the summarized within-study comparison (including the non-experimental arm from subsection 2.1) is given in tables 1 and 2 in Shadish et al. (2008a). The traditional analysis of covariance without interaction terms (see subsection 2.2.1)[11] performed best for the adjustment of the quasi-experimental results of the language training and reasonably well for the correction of the estimated average total effect of the math intervention. The only adjustment method resulting in a larger bias reduction for the language outcome was the inverse-propensity weighting approach with weights based on the propensity score (see paragraph 2.3.2) associated with the largest standard errors of all considered methods. The standard errors of the weighting approach are probably inflated by the presence of very extreme propensity scores (see Shadish et al., 2008a, and the discussion provided above). Otherwise, no significant difference between the adjustment methods could be found. Hence, the authors conclude that *"in general our results do not support the preferential use of propensity scores over ordinary linear regression"*. Pohl et al. (submitted) draw a similar conclusion: *"Assuming linearity and extrapolating does seem to be an appropriate assumption in applications like those incorporated into past within-study comparisons."* Nevertheless, as pointed out by R. J. Little et al. (2008), the differences between treated and untreated students in the within-study comparison by Shadish et al. (2008a) are generally too small compared with the sampling error, which clearly limits the generalizability of the results. However, based on a review of further within-study comparisons, Glazerman, Levy, and Myers (2003) also conclude that ANCOVA and propensity score methods hardly differ with respect to the bias reduction they achieve.

The cited within-study comparisons (see also T. D. Cook et al., 2008) give reason to believe in the potential of adjustment methods for the estimation of average total effects in quasi-experimental designs. Since assignment modeling based approaches are often found to be very similar to the results obtained by outcome modeling (see also Posner, Ash, Freund, Moskowitz, & Shwartz, 2001; Stürmer et al., 2006; Zanutto, 2006) we conclude that the choice of an adjustment method does not play the most important role for the estimation of causal effects from quasi-experimental designs, at least with respect to the bias reduction.

2.7 Summary and Conclusion

In the previous chapter we presented the theory of stochastic causality as a background for the estimation of average total effects in quasi-experimental designs. In this chapter we have provided an example for a quasi-experimental evaluation of two educational interventions. With this example in mind, we presented a survey of adjustment methods, which were mainly applied to the estimation of the average total effect for the given example. In particular, we introduced the most challenging alternatives to generalized analysis

[11] The method is called ANCOVA using observed covariates in Shadish et al. (2008a). Note that the authors applied a standard backward selection procedure of main effects for this analysis.

of covariance, the adjustment methods based on assignment modeling with estimated propensity scores. Afterwards, we focused on some of the common problems for the application of the adjustment methods and reviewed within-study comparisons as one strategy to judge the appropriateness of the estimation of average total effects in quasi-experimental designs.

The results showed that the concrete adjustment method (either an outcome modeling or an assignment modeling approach) is secondary with respect to the bias reduction achieved in the presented examples, and that adjustment methods in general can yield trustable average total effects if carefully applied. With respect to statistical inference about the estimated treatment effect, most of the adjustment methods based on assignment modeling ignore the uncertainty associated with propensity score estimation. Therefore, in line with Hill and Reiter (2006) we suggest a comparison of different standard errors for the estimated average total effect. In conclusion, it is maintained that the analysis of covariance is still one of the feasible adjustment methods, if unbiasedness of the covariate-treatment regression can be assumed. The same conclusion, but based on a comparison of different adjustment strategies for a constructed real data example, can be found in Schafer and Kang (2008). The authors claim that the performance of the analysis of covariance can be greatly enhanced by considering nonlinear trends, summaries of the propensity scores, baseline by treatment interactions and the usage of robust standard errors: *"With these enhancements, ANCOVA is no longer one of the worst ways to estimate an* ATE*, and it may be one of the best"* (p. 306).

In the next chapter we will focus mainly on two of the suggestions discussed by Schafer and Kang (2008) and we will derive structural equation models with nonlinear constraints to test hypotheses about average total effects. The purpose of this thesis can therefore be summarized as the application of structural equation modeling to obtain correct standard errors for an extended analysis of covariance with interaction terms and for heteroscedastic residual variances. The developed models can be extended further to include summaries of the propensity scores and nonlinear relationships, as we will describe in section 5.4.2, and the developed models — although discussed for the simple case with a single perfectly measured covariate — can be applied to account for measurement error on the covariates.

Chapter 3
Implementation of Generalized Analysis of Covariance

3.1 Introduction

In chapter 1, we summarized the theoretical background for the average total effect and highlighted the crucial points for valid causal inference in quasi-experimental studies. A survey of adjustment methods was presented in chapter two, illustrated with empirical examples. We are now discussing the implementation of the generalization of the analysis of covariance as suggested by Steyer et al. (in press) in more detail.

The term *analysis of covariance* is used for analysis of covariance without an interaction between treatment and covariates (for a review of different meanings of the term analysis of covariance see Cox & McCullagh, 1982). *Generalized analysis of covariance* is used in line with Steyer et al. (in press) as an adjustment method in the tradition of the analysis of covariance, applicable for quasi-experimental studies, i. e., flexible to account for heterogeneity of residual variances and to consider covariate-treatment interactions. We shall begin with a short discussion of these requirements for the implementation of a generalized analysis of covariance as implied by the theory of stochastic causality. The subsequent sections of this chapter will deal with the implementation in the two more general statistical frameworks: The general linear model in section 3.2 and the framework of structural equation modeling in section 3.3. Finally, in section 3.4, research questions are formulated, which are investigated by a Monte Carlo simulation and presented in chapter 4.

3.1.1 Covariate-Treatment Interactions

The substantive meaning of interactions between covariates and the treatment variable can be described as the ability of the regression model used to allow an adjustment for different regressive dependencies of the outcome variable and the covariate(s) for the different treatment conditions. Imagine, for example, that the selection into treatment groups in an educational setting, as described in the last chapter, is confounded due to students' baseline performance as measured by a pre-test. If, furthermore, the two treatment conditions differ with respect to the incorporation of students' knowledge in the instructional process, it is very likely that we have to take into account different dependencies between pre-treatment performance and the outcome across the two groups. According to these different dependencies, the individual treatment effects depend on the value(s) of the covariate(s).

The need to model an interaction between the covariates and the treatment variable in the covariate-treatment regression follows directly from the theory of stochastic causality, when average total effects are identified as the difference in the adjusted means (see subsection 1.1.8). As described in subsections 1.1.8 and 2.5.6, the assumption of Z-conditional unbiasedness of the covariate-treatment regression and the correct specification of the functional form of the covariate-treatment regression (when necessary) is of particular importance for the identification of average total effects. Therefore, although the inclusion of an interaction term seems a minor step in the direction of a more flexible covariate-treatment regression, group-specific covariate-regressions are an important implication from the theory of stochastic causality. For a linear parameterization of the *fundamental equation* introduced in Equation (1.26), equal regression coefficients in the group-specific regression of Y on Z are not necessary for the identification of the adjusted means. In other words, even a simple linear parameterized *effect function* $g_1(Z)$ with only one covariate Z can have a non-zero regression coefficient for the covariate. Making the assumption of *parallel regression slopes* in Equation (2.3) on page 27, that is assuming $\gamma_{11} = 0$, yields the well-known regression representation of the traditional analysis of covariance (as described in subsection 2.2.1).

Interaction terms for the analysis of covariance in randomized experiments are discussed, e. g., by Ganju (2004) and Moore and Laan (2009). In general, the analysis of covariance is recommended for experimental designs with randomization to increase the statistical power (see, e. g., Feldt, 1958, for an early discussion of covariate-treatment interactions), when stratification is not feasible. As derived by Yang and Tsiatis (2001) the analysis of covariance without interaction, as well as the paired t–test, are asymptotically less efficient than a generalized analysis of covariance including a covariate-treatment interaction (see also Tsiatis, Davidian, Zhang, & Lu, 2008).

Although T. D. Cook and Campbell (1979) have already presented an analysis of covariance with mean-centered covariates and interaction terms, some textbooks (still) argue that for non-parallel regression slopes, average effects are not appropriate. For instance, Kutner et al. (2005, p. 923, nomenclature changed) argue that *"when the treatments interact with the concomitant variable Z, resulting in nonparallel slopes, covariance analysis is not appropriate"*. They suggest the estimation and comparison of separate treatment regressions, but do not generalize their approach to the average total effects. As Rutherford (2001, p. 146) summarized, *"heterogeneous regression ANCOVA presents a problem of experimental effect determination and description"*. We will briefly discuss the conventional treatment of non-parallel regression slopes, i. e., methods known as *probing interactions* which were developed, for instance, by Rogosa (1980) in subsection 3.2.4. For a review of the typical mistreatment of ANCOVA, when interaction terms are present, see, for example, Engqvist (2005).

The incorporation of interaction terms in the covariate-treatment regressions is also advocated in the literature. For example, Gelman (2004) argues that an interaction between the treatment variable and covariates is a general phenomenon that can be seen as being derived from an underlying variance components model when individual treatment effects can vary with pretreatment covariates. Interaction effects can also be seen as a special case of nonlinear effects (see, e. g., Moosbrugger, Schermelleh-Engel, A., & Klein, in press). The inclusion of interaction terms, as well as the modeling of nonlinearities of the covariate-treatment regression, have recently also been suggested by Schafer and Kang (2008) to enhance the performance of the traditional analysis of covariances for non-randomized studies. Especially covariate-treatment interaction pose a basic challenge for the implementation of a generalized analysis of covariance when the focus lies on the average total effects. As we will point out, this is true for estimators developed within the general linear model as well as for estimators of the average total effect based on structural equation models with nonlinear constraints.

3.1.2 Heterogeneity of Residual Variances and Heteroscedasticity

In the theory of stochastic causality summarized in chapter 1 it is neither assumed that the true outcomes are fixed values for each subjects nor it is assumed that the difference between the true outcomes are the same for all subjects. Instead, variability for the true outcomes and heterogeneity of individual causal effects are incorporated in the definition of causal effects. Here, we pick up this issue again because of its importance for the development of appropriate models and statistical tests.

Heterogeneity of Between-Group Residual Variances In general, the concept of an individual treatment effect has already been introduced by Neyman (1923/1990). If individual treatment effects are taken into account, different implied residual variances for the treatment regression and the covariate-treatment regression conditional on the treatment variable are expected because the treatment will add additional variability to the true outcome. Randomization only guarantees equal pre-treatment variances. But even in the randomized case, the variances of the dependent variable will likely differ between groups after the treatment has been applied because the different treatment conditions might lead to different individual (total) effects (see for a similar argumentation, e. g., Blundell & Costa Dias, 2000, as well as Caliendo, 2006).

Formally, for instance, under the simplifying restriction $\mathfrak{D}_X = \sigma(U)$ we can consider the *unit-treatment regression* as

$$E(Y|X,\mathfrak{D}_X) \quad = \quad E(Y|X,U) \quad = \quad \tau_j + \delta_{jk} \cdot X, \tag{3.1}$$

with the residual $\varepsilon \equiv Y - E(Y|X,U)$.[1] We can decompose the variance of the outcome variable for the two treatment conditions $X = j$ and $X = k$ as

$$\begin{aligned} Var_{X=j}(Y) &= Var_{X=j}(\tau_k) + Var_{X=j}(\delta_{jk}) + 2Cov_{X=j}(\tau_j, \delta_{jk}) + Var_{X=j}(\varepsilon) \text{ and} \\ Var_{X=k}(Y) &= Var_{X=k}(\tau_k|Z) + Var_{X=k}(\varepsilon). \end{aligned} \tag{3.2}$$

Under the assumption of Z-conditional unbiasedness of the covariate-treatment regression $E(Y|X,Z)$ for a dichotomous X (see subsection 1.1.5), we can write this regression in a similar way as

$$E(Y|X,Z) = E(\tau_j|Z) + E(\delta_{jk}|Z) \cdot X, \tag{3.3}$$

with the residual $\varepsilon \equiv Y - E(Y|X,Z)$. If we now consider the conditional variances of Y given the treatment, we see that they are expected to differ:

$$\begin{aligned} Var_{X=j}(Y) &= Var_{X=j}(\tau_k|Z) + Var_{X=j}(\delta_{jk}|Z) + 2Cov_{X=j}(\tau_j, \delta_{jk}|Z) + Var_{X=j}(\varepsilon) \\ Var_{X=k}(Y) &= Var_{X=k}(\tau_k|Z) + Var_{X=k}(\varepsilon). \end{aligned} \tag{3.4}$$

Hence, the implied variances differ between groups at least by the conditional variance of the true total effect variable given the covariate(s) $Var_{X=j}\big(\delta_{jk}|Z\big)$. This difference should be expected, even if we take into account that $Var_{X=j}(\tau_k|Z) = Var_{X=k}(\tau_k|Z)$ and $Var_{X=j}(\varepsilon) = Var_{X=k}(\varepsilon)$, as a consequence of randomization.[2]

The substantive meaning of this theoretically implied difference of the outcome variable's variance between groups is that even though we can identify the average total effect as the difference between the adjusted means unbiasedly (provided that the covariate-treatment regression is Z-conditionally unbiased and that the functional form of the regression is precisely correctly modeled), we do not necessarily expect that the residuals for this covariate-treatment regression be (normally) distributed with the same variance. A similar formulation can be found in Bryck and Raudenbush (1988, p. 397) who pointed out that *"as we allow treatment effects to have a distribution, heterogeneity of variance across groups will occur"*. Within this thesis we use the term *heterogeneity of between-group residual variances* when referring to this phenomenon.

Heteroskedasticity Heterogeneity of between-group residual variances for the covariate-treatment regression is a special case of a more general phenomenon. As Hayes and Cai (2007) remark, *heteroskedastic-*

[1] For the subsequent consideration of the outcome variable's variance it is important to note that the residual ε is uncorrelated with the true outcome variable in the control group τ_j and with the true total effect variable δ_{jk} (see, e. g., Steyer, 2003).

[2] Within this thesis we do not consider the additional covariance between the true outcome variable in group $X = k$ and the true total effect variable, i. e., we simplify the considerations by assuming $Cov_{X=j}(\tau_k, \delta_{jk}|Z) = 0$.

ity[3] can result from a variety of different processes, for example, as the consequence of a misspecified functional form of a regression model. For instance, Angrist and Pischke (2009, p. 35) point out that for a linear ordinary least-squares regression $Q(Y|X)$ [i. e., a linear quasi-regression, see Steyer, 2003] heteroskedasticity seems natural when the linear ordinary least-squares regressions are considered as approximations of the true regressions $E(Y|X)$. For practical applications, the fit of parametric approximations might vary over the observed range of the regressors, resulting in the finding that even if the conditional variance of the outcome variable Y is constant, the residual variances increase with the square of the difference between the true regression $E(Y|X)$ and the least-squares estimated linear quasi-regression $Q(Y|X)$ [see also White, 1980b].

It is important to note that the underlying theory for generalized analysis of covariance does neither require that the heterogenous residual variances can be explained by the covariates nor is the heterogeneity predicted as a consequence of the approximating the functional form of the covariate-treatment regression, for which reason we differentiate *heterogeneity of between-group residual variances* [as a consequence of $Var(\delta_{10}) > 0$] from the more general *heteroskedasticity*. Nevertheless, for the statistical model both phenomena might invalidate common assumptions regarding the variance of the residuals for the regression of Y on X and Z. This might lead to biased standard error estimators within the general linear model. For the structural equation models discussed in this chapter, heterogeneity of between-group residuals can yield what is known as *specification error* and might affect the estimated parameters (and consequently the *ATE*–estimator as a nonlinear function of estimated parameters) as well as their standard errors.

3.1.3 Stochasticity of Regressors

Generalized analysis of covariance is studied within this thesis as an adjustment method for the estimation of average total effects for data obtained from quasi-experimental designs as defined in section 1. In those designs, the outcome variable Y, the covariates Z and the treatment variable X are *random variables* with a joint multivariate distribution. We made this conception explicit by describing the underlying structure of the quasi-experimental data (i. e., the sample) as obtained from repeated single-unit trials (see the description of the single-unit trial in section 1.1.9).[4]

The distinction between *fixed regressors* (*fixed-X assumption*, see, e. g., Maddala, 1992; Muller & Stewart, 2006), and *random regressors* (e. g., the random model mentioned by Mendoza & Stafford, 2001) is well

[3] We use the term heteroscedasticity here in line with the literature of the general linear model, e. g., Rao and Toutenburg (1999), that is we consider $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i^2) = \sigma_i^2$ for $i = j$ and $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$.

[4] A similar, informal description of the same conceptualisation can be found, for instance, in Mendoza and Stafford (2001, p. 651): A *"random model is more appropriate for nonexperimental situations in which the levels of the independent variables are not fixed a priori. It is common in education and other social sciences to have studies in which the levels of the multiple independent variables for each experimental unit cannot be controlled and are available only after the observations are made. This type of design clearly falls under the random model."*

known in the literature for regression models (see, e. g., Sampson, 1974).[5] Nevertheless, both regression models are often not handled separately. For instance, Rencher and Schaalje (2007) point out that there are a lot of similarities between the regression models with fixed regressors and the regression models with random regressors. For example, the confidence intervals for the regression coefficients and for linear functions of the regression coefficients derived under the fixed-X assumption are valid for stochastic regressors as well. Different treatments of the random regressor cases and the fixed regressor cases are typically applied only for the following exceptions within the general linear model: Power calculations and sample size considerations (see, for example, Glueck & Muller, 2003; Shieh, 2003, 2006), and to confidence intervals of the squared multiple correlation coefficient (Gatsonis & Sampson, 1989; Algina & Olejnik, 2000), and with respect to multicollinearity (Ayinde, 2007).

With respect to the analysis of covariance, Quinn and Keough (2002)[6] name three necessary conditions, which must be fulfilled for the analysis of covariance in the presence of stochastic regressors. The first assumption, *homogeneity of variance*, was discussed in the previous subsection. A second condition named by Quinn and Keough (2002) is that the range of the covariate must sufficiently overlap between the groups. We have already discussed this condition in subsection 2.5.3 as it protects causal inference (in particular outcome modeling based adjustment procedures) against unjustified extrapolation and is of importance especially with respect to the functional form assumption. Finally, a third condition is concerned with parallel regression slopes, i. e., the exclusion of interactions between X and Z, which was our first requirement for a generalized analysis of covariance (see section 3.1.1). Unfortunately, Quinn and Keough (2002) do not offer any advice on how to continue with stochastic regressors when one of the conditions is not fulfilled.

Furthermore, the stochasticity of regressors is sometimes discussed for regression models with interaction terms. The inappropriateness of the fixed-X assumption for moderated multiple regression (MMR) was claimed, e. g., by Fisicaro and Tisak (1994).[7] Although it is known that the assumption of fixed regressors is at least questionable when interactions are present, it makes the resulting statistical model much

[5] The same is also claimed for the analysis of covariance, e. g., in a comment by Bahpkar in the included discussion of Cox & McCullagh, 1982: *"a distinction should be made between the case in which the covariate Z is fixed and the case in which the covariate Z is essentially random"* (p. 555).

[6] Quinn & Keough, 2002, p. 349: *"The covariate X is assumed to be a fixed variable with no error associated with it. This is the standard fixed-X assumption of linear regression. This assumption is almost never valid for ANCOVA [...] because the covariate is usually a random variable, just like the response variable. As we pointed out [...], X being a random variable in regression analysis usually results in underestimation of the true regression slope. If the assumptions about homogeneity of variance, range of covariate values and parallel slopes hold, there is no reason to suspect that the underestimation of the true pooled within-groups regression coefficient between Y and X will vary between treatments. Therefore, tests of significance should still be reliable. We know of no extension of the Model II regression approach [that is regression model with stochastic regressors, ...] to ANCOVA."*

[7] Fisicaro and Tisak (1994, p. 34 f.) summarize that *"...a critical aspect of MMR, the stochastics (i. e., the statistical and probabilistic nature) of the technique, has been largely ignored. An examination of the stochastics reveals that MMR is an appropriate technique when values of predictors (including hypothesized moderators) are selected (i. e., predictors are fixed variables) but is not an appropriate technique when values of predictors (including hypothesized moderators) are obtained through some sampling procedure (i. e., predictors should be viewed as random variables) and the joint distribution is multivariate normal."*

easier because no distributional assumptions for the regressors are necessary. Nevertheless, for example, Maddala (1992, p. 103) concludes: *"To obtain any concrete results with stochastic regressors, we need to make some assumptions about the joint distribution of Y and Z."* This calls our attention to the possible limits of the general linear model for testing hypotheses about the average total effect.

3.1.4 Summary and Outline

In this section we described three requirements for the statistical implementation of generalized analysis of covariance. Obviously, the derivation of a valid standard error estimator for the average total effect is not straightforward when stochastic regressors are considered and interactions between covariate(s) and treatment are incorporated.

In subsection 3.2.5 we will provide a detailed derivation of the consequences of stochastic covariates for unconditional inference about the average total effect based on ordinary least-squares regression (i. e., the general linear model). As we will point out, the variance of the average total effect estimator differs from the variance obtained conditional on the covariates as a consequence of the interplay between covariate-treatment interactions and the stochasticity of the covariates.

This derivation will be performed without further distributional assumption for Y, X and Z. Even though this will not give us the opportunity to correct the bias of the standard error estimator for the average total effect obtained from ordinary least-squares regression (see Maddala, 1992, cited above), it will provide us with important insights into the performance of test statistics based on the general linear model and the general linear hypothesis. Following the advice found in the literature that a joint distributional assumption is necessary to incorporate the regressors' stochastic nature, we will then turn to structural equation models, where a distributional assumption is often made for the common maximum likelihood estimation. In the light of the requirements discussed in the previous three subsections, we will present and develop different implementations of generalized analysis of covariance. Note that heterogeneity of residual variances as a special form of heteroskedasticity is a distinct requirement for the implementation of generalized analysis of covariances. We will discuss heteroskedasticity consistent estimators for the general linear model and analyze the model implied variance-covariance matrix for the derived structural equation models.

3.2 General Linear Model

In the following we shall present a brief description of the general linear model and summarize the two assumptions central to demonstrate the inadequacy of the implementation of generalized analysis of covariance based on ordinary least-squares estimated covariate-treatment regressions. Additionally, we shall describe different methods of standard error correction of the general linear model. Based on a review of

the common methods for probing interaction terms in moderated multiple regressions (i. e., a discussion of inference based on the conditional variance of the effect function) we will discuss unconditional inference about the average total effect. We will show that the average total effect cannot be tested in the framework of the general linear model because the simplifying assumption of fixed regressors for the least-squares estimation is violated when covariates and the treatment variable interact and when covariates are stochastic regressors.

3.2.1 Introduction

The well-known *general linear model* (GLM) is defined by the following equation (see, e. g., Rao, 1973):

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.5}$$

where $\mathbf{y}$ is a $N \times 1$ vector of the univariate dependent variable y, $\mathbf{x}$ is the $N \times K$ design matrix for the $K-1$ regressors and the constant, $\boldsymbol{\beta}$ is the $K \times 1$ vector of regression coefficients for $\mathbf{x}$, and $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of the residuals. The matrix $\mathbf{x}$ is known as the *design matrix*, i. e., a matrix composed of values of the regressors (see the discussion below). *Ordinary least-squares* estimates of the regression coefficients for the general linear model are obtained as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{y}, \tag{3.6}$$

and an unbiased estimator for σ^2 based on $\hat{\boldsymbol{\varepsilon}} = \mathbf{x}\,\hat{\boldsymbol{\beta}} - \mathbf{y}$, i. e., for the variance of the residual ε_i as the elements of $\boldsymbol{\varepsilon}$ is $\widehat{\sigma}^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$.

The least-squares estimator can be derived as the solution of a least-squares minimization without additional assumptions about the residual terms (see, e. g., Steyer, 2003). Furthermore, if the residuals are assumed to be normally distributed [see Equation (3.7)], the maximum likelihood estimator is identical to the ordinary least-squares estimator (see, e. g., Rao & Toutenburg, 1999, for a detailed discussion).

3.2.2 Assumptions

3.2.2.1 Homoskedasticity

For the general linear model, the assumption of normally distributed residuals in Equation (3.5) with $E(\boldsymbol{\varepsilon}) = 0$ and $Var(\varepsilon_i) = \sigma^2$ is often added to derive properties of the estimated coefficients (see, e. g., Steyer, 2003):

$$\epsilon \sim N(0, \sigma^2\mathbf{I}) \tag{3.7}$$

The diagonal elements of the $N \times N$ covariance matrix $\Sigma_{\epsilon\epsilon} = \sigma^2\mathbf{I}$ are assumed to be equal, which is known as the assumption of *homoskedasticity* of the error variances. All off-diagonal elements in $\Sigma_{\epsilon\epsilon}$ are assumed to be equal to zero, i. e., the residuals are assumed to be uncorrelated. As described in detail in section 3.1.2, the *homoskedasticity* assumption is likely to be violated for applications of the general linear model for inferences about the estimated average total effect, because we expect *heterogeneity of between-group residual variances* as a consequence of individual treatment effects.

A noticeable amount of literature deals with the robustness of the general linear model against heteroskedasticity, and misleading interpretations of the assumption of homogeneity of residual variance can be found in the literature (see for a summary Bobko & Russell, 1990). It is known that even in the presence of heteroskedasticity, estimates of the regression coefficients obtained by ordinary least-squares remain unbiased (as they are derived without assumptions about the variances of the residuals ϵ_i, see, for example, Hayes & Cai, 2007). Furthermore, the F–test used in regression analysis and analysis of variance is known to be robust against heterogeneity of residual variance and non-normality, in particular when group sizes are equal (see, e. g., Berry, 1993, p. 78). Violation of homoskedasticity is critical with respect to the standard error of estimated regression coefficients when group sizes are unequal, a common situation for the analysis of unbalanced quasi-experimental designs (see, e. g., Ito, 1980).[8] Furthermore, there is a known effect of heteroskedasticity on the statistical power of detecting unequal slopes in analysis of covariance (see, for example, Alexander & DeShon, 1994). This effect also applies to the statistical power of the detection of moderation effects (e. g., Aguinis, Petersen, & Pierce, 1999).

We have identified three important approaches to account for heteroskedasticity: Weighed least-squares, the transformation of variables, and the calculation of robust standard errors. To correct the standard errors of ordinary least-squares estimates for heteroskedasticity, alternative estimation methods like generalized and weighted least-squares have been suggested. By the same derivation we used to demonstrate that the theory of stochastic causality implies heterogeneity of residual variances (see above), the structure of the variance-covariance matrix of residuals can be predicted and taken into account for the estimation process (see, for example, Cox & McCullagh, 1982). Weighted least-squares can be used to obtain an estimator which gives unbiased standard errors by specifying the structure of the residual matrix in terms of weights (see, e. g., Kutner et al., 2005, ch. 11, or S. Weisberg, 2005, for an introduction to weighted least-squares and Wilcox & Keselman, 2004, for a general summary of robust methods dealing with heterogeneity of residual variances, as well as Cai & Hayes, 2008, for an up-to-date overview of how to handle heteroskedasticity of an unknown form). We do not consider weighted least-squares within this thesis because we will focus on structural equation models under a multivariate normality assumption. Some

[8] Some authors suggest avoiding unbalanced designs by *throwing out data* (see, for instance, Scheiner & Gurevitch, 2001, p. 119).

authors suggest the transformation of the variables to handle heteroskedasticity of residual variances (e. g., Carroll & Ruppert, 1988). We do not follow this approach either because as Long and Ervin (2000) point out: If there are theoretical reasons to believe that errors are heteroscedastic around the correct functional form, transforming the dependent variable is inappropriate. Furthermore, it should at least be mentioned that a lot of diagnostic techniques are discussed in the literature for detection of heteroskedasticity (see Darken, 2004, for an overview, and, e. g., D. R. Cook & Weisberg, 1983, for a test statistic based on the score statistic).

Liang and Zeger (1986) suggest the application of robust standard errors based on White (1980a) to draw valid inference about the estimated parameters, even if the assumption of homoskedasticity is not fulfilled. In order to make the ordinary least-squares estimators (and the general linear model) more robust to the violation of assumptions, a plethora of different corrections have been developed. Two of them are described in more detail in the following paragraphs: Heteroskedasticity consistent estimators and adjusted standard errors for regression estimates.

Heteroskedasticity Consistent Estimators Standard errors for regression coefficients obtained from the general linear model are expected to be biased due to heterogeneity of residual variance (heteroskedasticity) for unequal group sizes. A very general approach to deal with this bias is the application of robust standard errors (developed by White, 1980a). These so-called *sandwich estimators* are very popular in economics (see, e. g., Kleiber & Zeileis, 2008) and nowadays implemented, for example, in various `R`–packages (e. g., Zeileis, 2006), or as an additional macro code (see, for example, Hayes & Cai, 2007). A detailed description of White's heteroscedasticity consistent estimator can be found, e. g., in Greene (2007, ch. 10). The procedures are based on the *heteroskedasticity consistent covariance matrix* estimation (HCCM); different versions of the correction exist (see J. G. MacKinnon & White, 1985, and Zeileis, 2004, for a summary of their properties and their implementation). In a Monte-Carlo simulation, Long and Ervin (2000) found that the `HC3` estimator should be applied to small samples $N < 25$. As Zeileis (2004) points out, the `HC4` estimator recently suggested by Cribari-Neto (2004) further improves the small sample performance, especially in the presence of influential observations. The correction can be applied to the centering approach as well as to the general linear hypothesis. This is possible because of the general nature of the sandwich estimator. In line with Zeileis (2004), we will study the performance of `HC3` and `HC4` in the simulation study presented in chapter 4.

Standard Error for Regression Estimate A very promising approach for the estimation of the average total effect based on predicted values was described by Schafer and Kang (2008). We have already described this approach in subsection 2.2.4. For the simple linear regression with one covariate, this approach is al-

gebraically equivalent to the sample estimator of the expectation of the effect function (for a group-specific specification of the covariate-treatment regression):

$$\begin{aligned}
\widehat{ATE}_{10} &= \tfrac{1}{N}\textstyle\sum_i \left[\hat{y}_{i1} - \hat{y}_{i0}\right] \\
&= \tfrac{1}{N}\textstyle\sum_i \left[\left(\hat{\beta}_{10} + \hat{\beta}_{11}\cdot z_i\right) - \left(\hat{\beta}_{00} + \hat{\beta}_{01}\cdot z_i\right)\right] \\
&= \tfrac{1}{N}\textstyle\sum_i \left[\left(\hat{\beta}_{10} - \hat{\beta}_{00}\right) + \left(\hat{\beta}_{11} - \hat{\beta}_{01}\right) z_i\right] \\
&= \left(\hat{\beta}_{10} - \hat{\beta}_{00}\right) + \left(\hat{\beta}_{11} - \hat{\beta}_{01}\right)\widehat{\mu}_Z .
\end{aligned} \tag{3.8}$$

Schafer and Kang (2008, p. 293) provide formulas for standard errors that do not assume correctly specified implied variances for the outcome model, and should therefore be robust with respect to heterogeneity of residual variances: *"Our standard errors are robust to misspecification of mean-variance relationships, whereas the so-called model-based standard errors typically provided by linear regression software are not."* Hence, we mention this approach in this subsection again. For the derivation of the standard errors for the regression estimate see Schafer and Kang (2008, p. 311).

We conclude that the general linear model without an additional adjustment for heteroskedasticity is not suitable for statistical inference about the average total effect when groups are of unequal size (see also Hartenstein, 2005). For equal group sizes, we expect the methods based on the general linear model to be robust against heteroskedasticity, at least for conditions without covariate-treatment interactions.

3.2.2.2 Fixed Regressors

Within the general linear model, the values of the design matrix are assumed to be fixed and non-stochastic quantities. This conflicts with the basic requirement for the implementation of a generalized analysis of covariance presented in section 3.1.3. If we consider the simple covariate-treatment regression again, for a univariate covariate Z with a linear parameterization of the intercept function and the effect function [see Equation (1.29) on page 17], we can write the interesting part of the design matrix $\mathbf{x} = \left(\mathbf{1}, \mathbf{X}_{fixed}\right)$ out as

$$\mathbf{X}_{fixed} = \begin{pmatrix} x_1 & ,\ldots, & x_N \\ z_1 & ,\ldots, & z_N \\ x_1 \cdot z_1 & ,\ldots, & x_N \cdot z_N \end{pmatrix}^T , \tag{3.9}$$

with x_i and z_i as known constants (fixed values of the regressors), and $x_i \cdot z_i$ as the simple products necessary to obtain a regression coefficient for the interaction term. Hence, with respect to the observed random variables opposed to the theory of stochastic causality only the elements of the vector $\mathbf{y}$ of the general linear model $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are assumed to change with repeated sampling.

Different arguments in support of this so-called *fixed-X assumption* for the traditional ANCOVA were summarized in section 3.1.3. Neter, Wasserman, and Kutner (1983, p. 83 f.) describe an interplay between the stochasticity of regressors and the homoskedasticity assumption mentioned above. They argue that as long as the conditional distributions of y_i given the regressors x_i are normal and independent with variance σ^2, and as long as the x_i are independent random variables whose probability distribution does not involve the regression parameters, the randomness of the regressors can be ignored (see also Ryan, 1996, p. 34). Flory (2004) presented results of a simulation study for a test of the hypothesis $ATE_{10} = 0$ within the general linear model for data generated with homogenous residual variances, equal group sizes and different interaction effects. For strong interaction effects he found heavily inflated type-I-error rate for a sample size of $N = 1000$. Hence, at least according to the results of this simulation study, the robustness of the linear regression model against violations of the fixed-X assumption does not hold for regression models with covariate-treatment interactions.

Although Henderson (1982) presents an approach to the analysis of covariance within the framework of *mixed models*, and Milliken and Johnson (2002) describe a *random effects model* with covariates considered as random within the framework of the *generalized linear mixed model* (GLMM), we are not aware of further approaches[9] dealing with the robustness of the general linear model with respect to violations of the fixed X-assumption.

3.2.3 General Linear Hypothesis

For a test of the average total effect estimated with the help of a linear parameterized covariate-treatment regression (see the presentation of the moderated regression in subsection 2.2.3), the general linear model was studied by Flory (2004) as mentioned in the previous subsection. The general linear model offers a flexible way of hypotheses testing for linear combinations of elements of β, known as the *(general) linear hypothesis* (see, for example, Steyer, 2003). The H_0 of the general linear hypothesis is defined as

$$H_0: \quad \mathbf{A\beta} - \boldsymbol{\delta} = \mathbf{0}. \tag{3.10}$$

Here $\boldsymbol{A}$ is a $M \times K$ matrix containing M linearly independent combinations of the K parameters of the model, $\boldsymbol{\delta}$ contains the hypothesized values of these contrasts, and $\boldsymbol{\beta}$ is the $K \times 1$ vector of the regression coefficients.

Under the null hypothesis and the assumptions described in section 3.2.2 it is possible to calculate a test statistic that follows a central F–distribution (Rao & Toutenburg, 1999). This F–test for the general linear hypothesis was applied by Flory (2004) for testing hypotheses about the average total effect. Accordingly,

[9] The only exception is an approach developed by Rogosa (1980), see paragraph 3.2.4 for a discussion of this statistical procedure.

the hypothesis was formulated in terms of the $\boldsymbol{A}$–matrix. For the simple model with one covariate and $\boldsymbol{\beta} = \left(\gamma_{00} \;\; \gamma_{01} \;\; \gamma_{10} \;\; \gamma_{11}\right)^T$, the corresponding $\boldsymbol{A}$–matrix for the hypothesis $H_0 : ATE_{10} = E\left(g_1(Z)\right)$ is

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 1 & E(Z) \end{pmatrix}. \tag{3.11}$$

The elements of $\mathbf{A}$ are assumed to be known (or fixed) numbers. The hypothesis tested by the general linear hypothesis is therefore assumed to be *linear* in their parameters (see, for example, Rutherford, 2001). To utilize the general linear hypothesis, the regression coefficients $\boldsymbol{\beta}$ must be replaced by the estimated coefficients $\hat{\boldsymbol{\beta}}$ from Equation (3.6). Unfortunately, the average total effect estimator is a nonlinear function of model parameters if the unknown true expectation of the covariate $E(Z)$ in Equation (3.11) is replaced by the estimated mean of the covariate $\widehat{\mu}_Z$. If the covariate is stochastic, $\widehat{\mu}_Z$ is an estimated parameter and the product of $\hat{\gamma}_{11} \cdot \widehat{\mu}_Z$ is part of the hypothesis. As a consequence, the hypothesis about the average total effect is nonlinear.

Predictive Simulations Gelman and Hill (2007, p. 180) suggest a simulation-based procedure for statistical inference in the case of nonlinear predictions (labeled as *predictive simulations*). We studied the approach as an alternative method in the conducted simulation study and would therefore like to give a brief description in this paragraph.

The aim of the predictive simulations is to obtain an alternative standard error for an average effect for the case of an outcome regression with interaction between covariate and treatment.[10] The suggested approach is to fit the ordinary least-squares moderated regression $E(Y|X,Z) = \left(\gamma_{00} + \gamma_{01} Z\right) + \left(\gamma_{10} + \gamma_{11} Z\right) \cdot X$, and to draw $N_{\text{Rep}} = 1000$ simulated regression parameters $\tilde{\gamma}$ out of an multivariate normal distribution. For large samples, the distribution of the regression coefficients is multivariate normal[11] (under the additional assumptions for statistical inference based on ordinary least-squares regressions, see above).

For each replication $r = 1, \ldots, N_{\text{Rep}}$ the average effect is computed as the sum over all units $i = 1, \ldots, N$ (compare to the *average distance* which will be discussed on page 64):

$$\widehat{ATE_{10}}_r = \sum_{i=1}^{N} \left(\tilde{\gamma}_{10_r} + \tilde{\gamma}_{11_r} \cdot z_i\right). \tag{3.12}$$

[10] The corresponding `R` - package is published on the authors' homepage: *http://www.stat.columbia.edu/ gelman/arm/*

[11] See, e. g., Bauer and Curran (2005, p. 377): *"from a frequentist perspective, the parameter estimates of the [moderated regression] model may be viewed as random variables characterized by a joint sampling distribution."* The parameters of this multivariate normal distribution are based on the ordinary least-squares estimated regression coefficients and their variances and covariances: $\tilde{\gamma} \sim MVN(1, \hat{\gamma}, \widehat{V}_{\hat{\gamma}} \cdot \sigma^2)$, with $\sigma = \hat{\sigma}\sqrt{(n-k)/X}$, $X \sim \chi^2(n-k)$, n as the degrees of freedom for the residual, and k as df for the model and $\widehat{V}_{\hat{\gamma}}$ as the estimated variance-covariance matrix of the regression parameters ($\widehat{V}_{\hat{\gamma}} = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$).

Finally, the average effect is computed as the mean over the simulated N_{Rep} average total effects estimated based on the simulated regression parameters

$$\widehat{\widetilde{ATE}}_{10} = Mean\left(\widehat{\widetilde{ATE}}_{10}\right) = \frac{1}{N_{\text{Rep}}} \sum_{r=1}^{N_{\text{Rep}}} \left(\widehat{\widetilde{ATE}}_{10_r}\right), \tag{3.13}$$

and an estimate of the simulation-based (adjusted) standard error is obtained as

$$\widehat{\text{S.E.}}(\widehat{\widetilde{ATE}}_{10}) = \sqrt{Var\left(\widehat{\widetilde{ATE}}_{10}\right)}. \tag{3.14}$$

The predictive simulation approach is founded on a Bayesian perspective, which we shall not further discuss here (see Gelman & Hill, 2007, and Gelman & Pardoe, 2007, for details). In light of the discussion of fixed versus stochastic regressors provided in this subsection, this approach is expected to fail, as the sampling variability of the values $Z = z$ of the covariate is not incorporated in the predictive simulation approach. We will highlight the similarity to the common methods for probing interactions in the next subsection.

3.2.4 Inference based on the Conditional Variance of the Effect Function

In the previous section we described the inappropriateness of the general linear hypothesis to test the average total effect. Following the concerns formulated by Crager (1987, p. 895), we will now present a discussion of inference based on the conditional variance of the effect function, i. e., conditional on the values of the covariate: *"An often-recommended solution to the problem of random covariates is to use fixed - covariate ANCOVA and to consider the analysis conditional on the set of observed covariate values. This leaves in doubt what the true significance level of the hypothesis tests would be with the sampling distribution of the covariate taken into account. Since the variance of ANCOVA parameter estimators is a function of the covariate values, there is cause for concern."*

As mentioned in the introduction of this chapter (see section 3.1), interactions are treated as a problem of *non-parallel regression slopes* in the ANCOVA literature. The typical treatment of non-parallel regression slopes is known as *probing interactions*, strongly inspired by the work of Rogosa (1980), who considered for a simple linear parameterization of the covariate-treatment regression [see Equation (2.3) on page 27] the difference between the sample regression lines as a function of the values of the covariate Z. With obtained estimates of the regression coefficients, the value of the *effect function* can be computed for each value z of the covariate Z as

$$\hat{g}_1(Z = z) \quad = \quad \hat{\gamma}_{10} + \hat{\gamma}_{11} \cdot z. \tag{3.15}$$

This point estimate is labeled as the *simple slope*, a linear combination of the coefficients estimated by ordinary least-squares from a (moderated) multiple regression model and the values of the covariates (Aiken & West, 1996, p. 25). Note that the values of the covariate are considered in Equation (3.15) as fixed and known, only the regression coefficients are treated as estimates (see fixed-X assumption in subsection 3.2.2.2).

Conditional Variance of the Effect Function and Center of Accuracy Following the idea of an *effect description* for a given value $Z = z$, Rogosa (1980) derived an estimator for the variance of the conditional effect as a transformation of the estimated variance of the regression coefficients $\widehat{Var}(\hat{\gamma}_{10})$, $\widehat{Var}(\hat{\gamma}_{11})$ and the covariance $\widehat{Cov}(\hat{\gamma}_{10}, \hat{\gamma}_{11})$. This variance of the effect function for a given value is due to the model-based variability of the estimated regression coefficients (see also Maxwell & Delaney, 2004):

$$\begin{aligned}
\widehat{Var}\left(\hat{g}_1(z)\right) &= \sigma^2\left[\tfrac{1}{n_A} + \tfrac{1}{n_B} + \tfrac{(z-\hat{\mu}_{Z|X=1})^2}{SSZ_{X=1}} + \tfrac{(z-\hat{\mu}_{Z|X=0})^2}{SSZ_{X=0}}\right] \\
&= \sigma^2\left[\tfrac{1}{n_A} + \tfrac{1}{n_B} + \tfrac{z^2}{SSZ_{X=1}} + \tfrac{\hat{\mu}^2_{Z|X=1}}{SSZ_{X=1}} - \tfrac{2z\hat{\mu}_{Z|X=1}}{SSZ_{X=1}} \quad + \quad \tfrac{z^2}{SSZ_{X=0}} + \tfrac{\hat{\mu}^2_{Z|X=0}}{SSZ_{X=0}} - \tfrac{z2\hat{\mu}_{Z|X=0}}{SSZ_{X=0}}\right] \\
&= \sigma^2\left[\tfrac{1}{n_A} + \tfrac{1}{n_B} + \tfrac{\hat{\mu}^2_{Z|X=1}}{SSZ_{X=1}} - \tfrac{z2\hat{\mu}_{Z|X=1}}{SSZ_{X=1}} \quad + \quad \tfrac{\hat{\mu}^2_{Z|X=1}}{SSZ_{X=0}} - \tfrac{2z\hat{\mu}_{Z|X=0}}{SSZ_{X=0}}\right] + z^2 \cdot \widehat{avar}(\hat{\gamma}_4) \\
&= \sigma^2\left[\tfrac{1}{n_A} + \tfrac{1}{n_B} + \tfrac{\hat{\mu}^2_{Z|X=1}}{SSZ_{X=1}} + \tfrac{\hat{\mu}^2_{Z|X=1}}{SSZ_{X=0}}\right] + 2z \cdot \widehat{Cov}(\hat{\gamma}_{10}, \hat{\gamma}_{11}) + z^2 \cdot \widehat{Var}(\hat{\gamma}_{11}) \\
&= \widehat{Var}(\hat{\gamma}_{10}) + 2z \cdot \widehat{Cov}(\hat{\gamma}_{10}, \hat{\gamma}_{11}) + z^2 \cdot \widehat{Var}(\hat{\gamma}_{11}).
\end{aligned} \tag{3.16}$$

This estimator of the variance of the effect function for a value $Z = z$ is also the starting point for the so-called *center of accuracy*. For a simple linear parameterized effect function with a single univariate covariate Z the center of accuracy is

$$\hat{z}_{CA} = -\frac{\widehat{Cov}(\hat{\gamma}_{10}, \hat{\gamma}_{11})}{\widehat{Var}(\hat{\gamma}_{10})}, \tag{3.17}$$

and in general, the center of accuracy can be obtained by differentiating the estimator of the variance for the conditional effects, i. e., $\widehat{Var}\left(\hat{g}_1(z)\right)$, with respect to the covariate and setting the resulting derivative equal to zero. For the value $\hat{z}_{CA}$ of the covariate Z obtained from Equation (3.17), Rogosa (1980) has shown that the (conditional) variance of the effect function is minimal by deriving the following decomposition:

$$\widehat{Var}\left(\hat{g}_1(Z = z)\right) = \widehat{Var}\left(\hat{g}_1(z_{CA})\right) + \widehat{Var}(\hat{\gamma}_{11}) \cdot (z - \hat{z}_{CA})^2. \tag{3.18}$$

Rogosa (1980) also noted that in case of an interaction, the classical ANCOVA, which assumes parallel regression lines, compares conditional means at the value $Z = \hat{z}_{CA}$ of the covariate, which are weighted averages across the groups (see also R. J. A. Little et al., 2000). The vertical distance between the two non-parallel regression lines at $Z = \hat{z}_{CA}$ equals the estimated difference of the adjusted means for an ANCOVA model with a common slope. In other words, if the interaction term is omitted although the regression slopes are not parallel, the (conditional) treatment effect for the value $Z = \hat{z}_{CA}$ of the covariate is estimated.

Average Distance and Average Total Effect In line with Rubin (1977), who defined the *average distance* as *"the sum of the vertical difference of the population within-group regressions weighted by the population distribution of Z"*, Rogosa (1980) gave an equivalent formulation for a covariate-treatment regression as moderated regression model:

$$\Delta(Z) = \frac{1}{N}\sum_{i=1}^{N}\left(\gamma_{10} + \gamma_{11}\cdot z_i\right). \tag{3.19}$$

This average distance yields an unbiased estimator of the average total effect (for instance, based on least-squares estimated regression coefficients) provided that the covariate-treatment regression is unbiased and the linear parameterization of the covariate-treatment regression holds (with $\hat{\mu}_Z = \frac{1}{N}\sum_{i=1}^{N} z_i$ as estimated mean of the covariate):

$$\widehat{ATE}_{10} = \frac{1}{N}\sum_{i=1}^{N}\left(g_1(Z = z_i)\right) = \hat{\gamma}_{10} + \hat{\gamma}_{11}\cdot\hat{\mu}_Z = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\gamma}_{10} + \hat{\gamma}_{11}\cdot z_i\right). \tag{3.20}$$

In other words, as described also by Finney, Mitchell, Cronkite, and Moos (1984), the average total effect in Equation (3.20) is estimated as the average of the simple slopes across all units (see also Aiken & West, 1996). We describe this relationship here in great detail to point out that the average total effect, although defined in a very general manner by Steyer et al. (in press), results in a quantity well known in the literature on the traditional treatment of the non-parallel regression slopes (Rogosa, 1980). Nevertheless, to obtain a valid standard error for the estimator in Equation (3.20) the properties of the ordinary least-squares estimation must be taken into account. This next logical step was not carried out by Rogosa (1980, p. 308, nomenclature changed): *"The statistical methods considered in this article condition on the values of Z that are observed, as is conventional in regression analysis. The data can be thought of as consisting of N Z, Y pairs that are a random sample from the population bivariate distribution of Z and Y. The values of Z are not chosen or fixed in advance. However, inferences from these data are restricted to sub populations having the same values or configuration of Z because inferences from the linear model are conditional on the observed values of Z."* Because of the product of the regression coefficient $\hat{\gamma}_4$ and the estimated mean of the covariate $\hat{\mu}_Z$ the estimator in Equation (3.20) is clearly not *conditional* on the covariate (see also subsection 3.2.3). This can also be verified simply by re-writing the definition of the average total effect for a discrete covariate Z in the following way:

$$ATE_{10} = E\left(g_1(Z)\right) = \sum_{z} g_1(Z = z)P(Z = z). \tag{3.21}$$

Obviously, this formulation of the expectation of the effect function (or the equivalent notation as integral over the distribution of a continuous covariate) incorporates the unconditional distribution of the covariate Z.

Note that Rogosa (1980) also developed a modified test statistic based on the adjusted mean difference for a classical ANCOVA model without interaction terms, which is proposed to be at least less influenced by violations of the assumption of equal regression slopes. Among others, Harwell and Serlin (1988) reported that this test statistic performs poorly in simulation studies (see for a summary Harwell, 2003). Therefore, we will not consider such an approach in this thesis.

In the remaining part of this subsection we present two of the traditional methods for probing interaction terms in order to show that all these methods are valid only for observed values of $Z = z$, i. e., under the fixed-X assumption. None of these techniques can be applied to derive valid unconditional inference about the average total effect.

Pick-a-point Technique The *pick-a-point* approach (Rogosa, 1980) involves plotting and testing the conditional effects at selected and meaningful values of the covariate (e. g., high, medium, and low). For ordinary least-squares estimates of the regression parameters and their variances and covariance, tests at particular values $Z = z$ can be performed with the help of Equation (3.18) [see, for example, Rogosa, 1980; Cohen et al., 2003; Bauer & Curran, 2005]:

$$t = \frac{\hat{g}_1(Z = z)}{\left[\widehat{Var}\left(\hat{g}_1(Z = z)\right)\right]^{1/2}}. \tag{3.22}$$

The application of the pick-a-point approach is suggested when an interaction between covariates and treatment is significant and if it makes substantive sense to compare groups at a particular value z of the covariate Z. The confidence interval for the comparison at $Z = z$ is: $CI_{Z=z} = \hat{g}_1(Z = z) \pm t_{\text{Crit}}\left[\widehat{Var}\left(\hat{g}_1(Z = z)\right)\right]^{1/2}$ (see, e. g., Jaccard, Turrisi, & Wan, 1990). While the application of the pick-a-point approach is simple due to available software implementations (see, e. g., O'Connor, 1988), typically only a small number of arbitrarily selected values are evaluated. Additionally, the selected z-values may even reside outside of the observed range of the regressor (e. g., one standard deviation above the mean of a skewed covariate, see Bauer & Curran, 2005). It might, therefore, be more interesting to examine the full range of the covariate as described in the next paragraph.

Johnson-Neyman Technique An alternative approach is the evaluation of the effect function $\hat{g}_1(Z = z)$ at each level of the covariate with the goal of determining regions of statistically significant (conditional) differences. The main idea behind this so-called *Johnson-Neyman Technique* (J-N) can be described as a reversion of the t-test presented in Equation (3.22). Now, the critical t_{Crit}-value (i. e., ± 1.96 for sufficiently large samples) is taken as a known constant, and the values $z^*_{1/2}$ of the covariate Z which solve the equation are searched for:

$$0 \quad = \quad t^2_{\text{Crit}} \cdot \left[\widehat{Var}\left(\hat{g}_1(Z = z)\right)\right] \quad - \quad \hat{g}_1(Z = z)^2. \tag{3.23}$$

The quadratic Equation (3.23) can be solved by plugging in Equation (3.15) and Equation (3.18), which leads to two real valued roots $z^*_{1/2}$ under certain conditions (see Rogosa, 1981, for a discussion). Bauer and Curran (2005) described a step-by-step derivation of the formula for the case of a simple linear parameterized effect function. The obtained two values $z^*_{1/2}$ are the boundaries of the so-called *region of significance*.[12]

Potthoff (1964) extended the J-N procedure by controlling the type-I-error rate for a large number of possible values of the covariate $Z = z$, at which a test can be performed. Larholt and Sampson (1995) analyzed the robustness of the regions of significance with respect to heterogeneity of residual variances for equal and unequal sample sizes. Furthermore, generalizations of this approach to a J-N type procedure for hierarchical linear models are given by various authors (see, for example, Tate, 2004; Miyazaki & Maier, 2005; Bauer & Curran, 2005).

The Johnson-Neyman technique as a method for *probing interaction* terms is also easily accessible due to various software implementations (see, for example, Karpman, 1983, for a SPSS implementation, Hunka & Leighton, 1997, for a generalization to more complicated effect functions in Mathematica, and Preacher, Curran, & Bauer, 2006, for a collection of online resources that implement these techniques for different models). A closely related strategy for describing the conditional effects is the computation of *confidence bands*, that is the computation or plotting of the confidence interval for the whole range of the covariate:

$$CB(z) \quad = \quad \hat{g}_1(Z = z) \pm t_{\text{Crit}} \left[\widehat{Var} \left(\hat{g}_1(Z = z) \right) \right]^{1/2}. \tag{3.24}$$

These confidence bands are narrowest at the value $Z = \hat{z}_{CA}$ because of the property presented in Equation (3.18).

Summary The common strategies for probing interaction terms were described in this subsection in more detail in order to show that they share the same idea of making inference about conditional effects given a fixed value z of the covariate Z. All methods rely on the estimated variance of the simple slope for a fixed value of the covariate $\widehat{Var}\left(\hat{g}_1(Z = z)\right)$ from Equation (3.16). Nevertheless, Rogosa (1980) was already aware of the average distance (in the tradition of Rubin, 1977), and he formulated an appropriate estimator for the linear parameterized analysis of covariance with interaction terms.

In the next subsection we shall present an analytical proof that the unconditional variance of the *ATE*–estimator is underestimated by the estimated variance obtained from the general linear model (i. e., from the ordinary least-squares regression).

[12] The same difficulties as for the pick-a-point technique with respect to the interpretation of the obtained boundaries of the region of significance occur when $z^*_{1/2}$ are outside of the observed range of the regressor Z.

3.2.5 Unconditional Inference about the Average Total Effect

We now turn to the consequences of the stocasticity of covariates Z for the estimation and testing of average total effects within the general linear model. We shall start with a discussion of the appropriateness of unconditional inference about the average total effect from moderated regression models, i. e., for a covariate-treatment regression with interaction term. This will be followed by a similar discussion for mean-centered covariates.

Moderated Regression In the following theoretical part of this thesis we will show that the average total effect cannot be tested in the framework of the general linear model because the simplifying assumption of fixed regressors for the least-squares estimation is violated when covariates and the treatment variable interact and when covariates are stochastic regressors. Therefore, we assume that although X and Z are stochastic regressors, the estimation of the covariate-treatment regression is obtained as described in the section 3.2.1, i. e., based on ordinary least-squares within the general linear model. Nevertheless, in contrast to the assumption $\mathbf{x} = \left(\mathbf{1}, \mathbf{X}_{fixed}\right)$ underlying the statistical model for the implementation of the regression $E(\mathbf{y}) = \mathbf{x\beta}$ within the general linear model [see Equation (3.9) on page 59], we define an alternative matrix $\mathbf{x}^{\sim} = \left(\mathbf{1}, \mathbf{X}_{RND}\right)$ with stochastic regressors $\mathbf{X}_{RND}$ defined as:

$$\mathbf{X}_{RND} = \begin{pmatrix} X_1 & ,\ldots, & X_N \\ Z_1 & ,\ldots, & Z_N \\ X_1 \cdot Z_1 & ,\ldots, & X_N \cdot Z_N, \end{pmatrix}^{\mathbf{T}}, \tag{3.25}$$

with Z_i and X_i as random variables (stochastic regressors), and $Z_i \cdot X_i$ as their product (covariate-treatment interaction). Based on the estimated regression coefficients

$$\hat{\boldsymbol{\beta}} = \left(\hat{\gamma}_{00} \;\; \hat{\gamma}_{01} \;\; \hat{\gamma}_{10} \;\; \hat{\gamma}_{11}\right)^{T}, \tag{3.26}$$

the *ATE*–estimator itself can be formulated again as the difference of the adjusted means, expressed as the nonlinear function of the estimated regression coefficients and the estimated unconditional mean of the covariate $\widehat{\mu}_Z$:

$$\widehat{ATE}_{10} = \hat{\gamma}_{10} + \hat{\gamma}_{11} \cdot \widehat{\mu}_Z. \tag{3.27}$$

Statistical inferences for ordinary least-squares estimated parameters are developed conditionally on the regressors $\mathbf{X}_{fixed}$ (see above). For a valid unconditional interpretation of the *ATE*–estimator with stochastic regressors it is necessary that the regression of the average total effect estimator on the stochastic regressors

$\mathbf{X}_{RND}$ equals the (unconditional) expectation of the average total effect estimator (estimated by ordinary least-squares conditional on the fixed regressors $\mathbf{X}_{fixed}$). Hence, we consider the regression $E(\widehat{ATE}_{10}|\mathbf{X}_{RND})$ analytically in order to prove conditional and marginal unbiasedness of the *ATE*–estimator for the covariate-treatment regression with interaction term:[13]

$$\begin{aligned} E(\widehat{ATE}_{10}|\mathbf{X}_{RND}) &= E\left[\hat{\gamma}_{10} + \hat{\gamma}_{11}\cdot\widehat{\mu}_Z|\mathbf{X}_{RND}\right] \\ &= E\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) + E\left[\hat{\gamma}_{11}\cdot\widehat{\mu}_Z|\mathbf{X}_{RND}\right] \qquad \text{because } \widehat{\mu}_Z = f(Z) = f(\mathbf{X}_{RND}) \\ &= E\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) + E\left(\hat{\gamma}_{11}|\mathbf{X}_{RND}\right)\cdot\widehat{\mu}_Z. \end{aligned} \tag{3.28}$$

As given, e. g., in S. Weisberg (2005, Equation A.22), the conditional estimators of the regression coefficients, i. e., $E\left(\hat{\boldsymbol{\beta}}|\mathbf{X}_{RND}\right)$ obtained by ordinary least-squares, are unbiased estimators for the regression coefficients $\boldsymbol{\beta}$, under the condition that the true mean function is fitted.[14] For the estimation of average total effects this condition is implied by the assumption of Z-conditional unbiasedness of the covariate-treatment regression with an appropriate functional form assumption.

If we make use of this unbiasedness property of the regression coefficients $\hat{\gamma}_{ij}$ as elements of $\hat{\boldsymbol{\beta}}$ as defined in Equation (3.26), we can continue and show that the conditional expectation of the *ATE*–estimator given the stochastic regressors, i. e., the regression of $\widehat{ATE}_{10}$ on $\mathbf{X}_{RND}$, equals the population value we are interested in:[15]

$$\begin{aligned} E(\widehat{ATE}_{10}|\mathbf{X}_{RND}) &= E\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) + E\left(\hat{\gamma}_{11}|\mathbf{X}_{RND}\right)\cdot\widehat{\mu}_Z \\ &= \gamma_{10} + \gamma_{11}\cdot\widehat{\mu}_Z \\ &= \gamma_{10} + \gamma_{11}\cdot\widehat{\mu}_Z + \gamma_{11}\cdot E(Z) - \gamma_{11}\cdot E(Z) \\ &= \gamma_{10} + \gamma_{11}\cdot E(Z) + \left[\gamma_{11}\cdot\widehat{\mu}_Z - \gamma_{11}\cdot E(Z)\right] \\ &= ATE_{10} + \left[\gamma_{11}\cdot\left(\widehat{\mu}_Z - E(Z)\right)\right]. \end{aligned} \tag{3.29}$$

According to Equation (3.29), the average total effect is estimated unbiasedly, if $\widehat{\mu}_Z = E(Z)$ [Z as fixed regressor], or $\gamma_{11} = 0$ (parallel regression slopes). Furthermore, the equality of $E(\widehat{ATE}_{10}|\mathbf{X}_{RND})$ and ATE_{10} is fulfilled for large sample sizes N, because $\lim_{N\to\infty}\left(\widehat{\mu}_Z - E(Z)\right) = 0$, i. e., if $\widehat{\mu}_Z$ gets close to the population value $E(Z)$. Therefore, the estimator of the average total effect based on ordinary least-squares is *consistent* and *unbiased*.

[13] Conditional unbiasedness means $E(\widehat{ATE}_{10}|\mathbf{X}_{RND}) = E(\widehat{ATE}_{10})$ and marginal unbiasedness means $E(\widehat{ATE}_{10}) = ATE_{10}$ (see, e. g., Senn et al., 2007, p. 5535).

[14] Unfortunately, the same is not true for the standard error of the regression coefficients. The conditional variances of the estimated regression coefficients [see Equation (3.6)] are not equal to the unconditional variances, because $Var\left(\hat{\boldsymbol{\beta}}|\mathbf{X}_{RND}\right) = \sigma^2\left(\mathbf{X}_{RND}{}^T\mathbf{X}_{RND}\right)^{-1}$, i. e., the variance of the estimated regression coefficients are determined by σ^2 and $\mathbf{X}_{RND}$ (see, e. g., Steyer, 2003, Equation 14.28). Furthermore, note that this also shows that the ordinary least-squares estimator for the average total effect $\hat{\gamma}_{10} = \widehat{ATE}_{10}$ for a covariate-treatment regression without interaction term [see Equation (2.1) on page 25] is conditionally and marginally unbiased.

[15] Note that we applied a "little trick" by adding and subtracting $\gamma_{11}\cdot E(Z)$ in the second step of Equation (3.29).

So far, we considered the estimator of the average total effect. Following Allison (1995) and Chen (2006), we are now decomposing the variance of the estimator for the average total effect. The unconditional variance is of interest, because the standard error of the *ATE*–estimator is obtained from the square root of the estimators' variance. In general, the relationship between the unconditional variance and the conditional variance is described by the *conditional variance identity* (see, e. g., Steyer, 2003; Wasserman, 2004):

$$Var\left(\widehat{ATE}_{10}\right) = E\left(Var\left(\widehat{ATE}_{10}|\mathbf{X}_{RND}\right)\right) + Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right). \tag{3.30}$$

The unconditional variance of the average total effect estimator is a composite of the expectation of the conditional variance $E\left(Var\left(\widehat{ATE}_{10}|\mathbf{X}_{RND}\right)\right)$, and the variance of the regression $Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right)$. The first summand of Equation (3.30) can be formulated as the expectation of the variance of the effect function for $Z = \widehat{\mu}_Z$ as developed by Rogosa (1980), [see Equation (3.16)]:

$$\begin{aligned} E\left[Var\left(\widehat{ATE}_{10}|\mathbf{X}_{RND}\right)\right] &= E\left[Var\left(\hat{\gamma}_{10} + \hat{\gamma}_{11}\cdot\widehat{\mu}_Z|\mathbf{X}_{RND}\right)\right] \\ &= E\left[Var\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) + Var\left(\hat{\gamma}_{11}\cdot\widehat{\mu}_Z|\mathbf{X}_{RND}\right) + 2\,Cov\left(\hat{\gamma}_{10}, \widehat{\mu}_Z\hat{\gamma}_{11}|\mathbf{X}_{RND}\right)\right] \\ &= E\left[Var\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) + \widehat{\mu}_Z^2\,Var\left(\hat{\gamma}_{11}|\mathbf{X}_{RND}\right) + 2\widehat{\mu}_Z\,Cov\left(\hat{\gamma}_{10}, \hat{\gamma}_{11}|\mathbf{X}_{RND}\right)\right]. \end{aligned} \tag{3.31}$$

The more interesting second summand in Equation (3.30) can be concretized by inserting Equation (3.29). Because neither the (true) average total effect ATE_{10}, nor the (true) expectation of the covariate $E(Z)$ have variance (per definition), this term can be reduced further:

$$\begin{aligned} Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right) &= Var\left(ATE_{10} + \left[\left(\gamma_{11}\right)\cdot\left(\widehat{\mu}_Z - E(Z)\right)\right]\right) \\ &= Var\left(\gamma_{11}\cdot\left(\widehat{\mu}_Z - E(Z)\right)\right) \\ &= \gamma_{11}^2\cdot Var\left(\widehat{\mu}_Z\right). \end{aligned} \tag{3.32}$$

This demonstrates that the unconditional variance of the *ATE*–estimator [$Var\left(\widehat{ATE}_{10}\right)$, see Equation (3.30)] is larger than Rogosa's variance for the average distance derived under the assumption of fixed regressors [see Equation (3.31)] under the following condition: The regression lines are not parallel ($\gamma_{11} \neq 0$, covariate-treatment interaction), and the estimator of the mean of the covariate has variance [$Var\left(\widehat{\mu}_Z\right) \neq 0$, Z as stochastic regressor]. Under the special condition that both parts are different from zero, the variance of the average total effect estimator (constructed from the ordinary least-squares estimates) underestimates the true variability of the corresponding *ATE*–estimator, i. e., the variance of the regression of the *ATE*–estimator on the stochastic regressors $\mathbf{X}_{RND}$ is different from zero and consequently the expectation of the conditional variance $E\left[Var\left(\widehat{ATE}_{10}|\mathbf{X}_{RND}\right)\right]$ underestimates the *ATE*–estimator's standard error.

The derivation presented in Equation (3.32) explains an important inconsistency in the literature: We now see that the statistical inference based on ordinary least-squares regression for the analysis of covariance model assuming fixed regressors is valid for the classical ANCOVA without interaction because from $\gamma_{11} = 0$ it follows that $Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right) = 0$. This finding reconfirms the many publications concluding that even if the assumption of fixed regressors is not reasonable, the inferences under this simple fixed model are valid for stochastic regressors.[16] Furthermore, our finding extends the conclusion that Gatsonis and Sampson (1989) have drawn for power and sample size calculations with stochastic regressors to generalized analysis of covariance, i. e., a distinction between random and fixed regressors in regression models is relevant for testing hypotheses about average total effects in quasi-experimental designs.

Mean-Centered Covariates For applied regression modeling with interaction terms, covariates or predictors are often mean-centered (Aiken & West, 1996, see also Jaccard et al., 1990; West & Aiken, 2005; Rencher & Schaalje, 2007; Schafer & Kang, 2008). Mean-centering is performed basically for two reasons: At first to reduce nonessential multicollinearity and thus to reduce instability of estimated regression coefficients and standard errors (Marquardt, 1980; T. D. Little, Bovaird, & Widaman, 2006).[17] Secondly, mean-centering is applied for an easier interpretation of the regression coefficients (see, for example, Aiken & West, 1996, Wainer, 2000, Kraemer & Blasey, 2004, and Angrist & Pischke, 2009 for an application of mean-centering for the interpretation of parameters from regression discontinuity designs). Note that mean-centering was also criticized as a technique for the estimation of interaction terms in ordinary least-squares regressions.[18] For a detailed technical description of the effect of mean-centering we refer to Draper and Smith (1998).

Yang and Tsiatis (2001) suggested for the estimation of average total effects a covariance analysis with interaction term based on mean-centered covariates and applied the method (labeled as *ANCOVA II*) to randomized pretest-posttest trials (see also Wooldridge, 2001). As we have already mentioned in the introduction (see subsection 2.2.3), for a linear parameterized covariate-treatment regression it follows from the definition of the average total effect that if the expectation of the covariate $E(Z)$ is zero, the average total effect reduces to a linear function of the regression coefficients (for the group-specific linear covariate-treatment-regressions), or equals the simple regression coefficient (for the single group formulation of the

[16] See, for example, Jaccard et al. (1990, p. 8): *"Some of the assumptions can be relaxed with minor consequences to inferential tests or parameter estimation, whereas other assumption violations are problematic. For example, although many research applications rely on cases where the predictors are stochastic rather than fixed, OLS remains a viable approach if one assumes stochastic predictors conditional on the actual sample of observed Z's."*

[17] Comments are given in the literature that mean-centering masks in fact collinearity and might be problematic for the diagnostics of essential multicollinearity (see, e. g., Belsley, 1984).

[18] See, for example, Cohen (1978, p. 866): *"Since this [...] is an algebraic necessity, it follows that whether independent variables are interval, ratio, ordinal, dichotomous, or nominal, whether or not they are scaled to zero means, standardized, or otherwise linearly transformed, whether or not interacting variables are orthogonal, whether they are from observational or experimental research, and whether single variables or sets of variables are at issue, partialed products are interactions and partialed powers are curve components."* A similar comment is given by Kromrey and Foster-Johnson (1998, p. 65): *"Little is gained, and much energy is lost, by devoting attention to methodological choices that appear different on the surface but are actually identical. Such is the case with mean centering in least squares regression analysis."*

same covariate-treatment regression). Hence, if $E(Z) = 0$, the estimator of the average total effect $\widehat{ATE}_{10}$ also reduces to $\hat{\gamma}_{10}$, and although this regression coefficient is estimated under the fixed-X assumption by ordinary least-squares, we know that the unconditional variance for $\hat{\gamma}_{10}$ is valid unconditionally.[19]

Flory (2008) studied the performance of mean-centering for the analysis of average (total) effects. In his simulation study he distinguished the case of mean-centering with the (assumed to be known) expectation of the covariate, $E(Z)$, from mean-centering based on the estimated mean of the covariate, $\widehat{\mu}_Z$. He discovered that mean-centering with the estimated mean of the covariate fails to take the stochasticity of the covariate appropriately into account. Based on the results obtained for the validity of the variance of the average total effect estimator presented in this subsection, we can now analytically explain his findings. For a covariate Z^*, mean-centered with the estimated mean $\widehat{\mu}_Z$, the average total effect can be rewritten as:

$$\begin{aligned} ATE_{10} &= E\left(g_1(Z^*)\right) && \text{with } Z^* := Z - \widehat{\mu}_Z \\ &= \gamma_{10} + \gamma_{11} \cdot E(Z^*) && \text{because } E(Z^*) = E\left(Z - \widehat{\mu}_Z\right) = E(Z) - \widehat{\mu}_Z \\ &= \gamma_{10} + \gamma_{11}\left(E(Z) - \widehat{\mu}_Z\right). \end{aligned} \tag{3.33}$$

The sample estimate of this average total effect is obtained as $\widehat{ATE}_{10} = \hat{\gamma}_{10} + \hat{\gamma}_{11} \cdot (\widehat{\mu}_Z - \widehat{\mu}_Z) = \hat{\gamma}_{10}$, because $\widehat{\mu}_Z$ is the sample estimate of $E(Z)$. As Aiken and West (1996, p. 38) argue for the moderated regression, the average over the simple slopes [see Equation (3.20) in section 3.2.4] reduces to the value of the simple regression coefficient [$\hat{\gamma}_{11}$ in Equation (3.20) on page 64]. To judge the appropriateness of the variance of the estimator obtained by ordinary least-squares, i. e., conditional on $\mathbf{X}_{RND}$, we analyze the regression of the *ATE*–estimator on the stochastic regressors $\mathbf{X}_{RND}$ for the mean-centered covariate again:

$$\begin{aligned} E(\widehat{ATE}_{10}|\mathbf{X}_{RND}) &= E\left(\hat{\gamma}_{10}|\mathbf{X}_{RND}\right) \\ &= \gamma_{10} \\ &= \gamma_{10} + \gamma_{11} \cdot E(Z^*) - \beta_{11} \cdot E(Z^*) \\ &= ATE_{10} - \gamma_{11} \cdot E(Z^*) \\ &= ATE_{10} - \gamma_{11} \cdot \left(E(Z) - \widehat{\mu}_Z\right). \end{aligned} \tag{3.34}$$

For the derivation in Equation (3.34), we inserted the definition of the average total effect after adding and subtracting $\gamma_{11} \cdot E(Z^*)$. According to the decomposition we introduced in section 3.1.3 [see Equation (3.30)

[19] This can be verified by reproducing the argumentation given in subsection 3.1.3 The average total effect is estimated as $\widehat{ATE}_{10} = \hat{\gamma}_{10}$ and therefore the estimator is unbiased because from $E\left(\hat{\beta}_{ij}|\mathbf{X}_{RND}\right) = \beta_{ij}$ (see, e. g., S. Weisberg, 2005) it follows that $E(\widehat{ATE}_{10}|\mathbf{X}_{RND}) = ATE_{10}$. If we now consider the decomposition $Var\,(\widehat{ATE}_{10}) = E\left(Var\,(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right) + Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right)$ again, the second part vanishes as $Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right) = Var\,(ATE_{10}) = 0$, see Equation (3.25) for the definition of $\mathbf{X}_{RND}$, and also Crager (1987) for a similar conclusion.

on page 69], we finally consider the variance of the regression considered in Equation (3.34) of the ordinary least-squares estimated average total effect on the stochastic regressors $\mathbf{X}_{RND}$ again:

$$\begin{aligned} Var\left(E(\widehat{ATE}_{10}|\mathbf{X}_{RND})\right) &= Var\left(ATE_{10} - \left[\gamma_{11}\cdot\left(E(Z) - \widehat{\mu}_Z\right)\right]\right) \\ &= Var\left(ATE_{10} - \gamma_{11}\cdot E(Z) - \gamma_{11}\cdot\widehat{\mu}_Z\right) \\ &= \gamma_{11}^2\cdot Var\left(\widehat{\mu}_Z\right). \end{aligned} \tag{3.35}$$

The comparison of Equation (3.32) and Equation (3.35) reveals that mean-centering (with the estimated mean of the covariate) does not abolish the underestimation of the variance of the *ATE*–estimator caused by the stochasticity of the covariate.

3.2.6 Summary

For covariate-treatment regressions with interaction terms, the unconditional variance of the *ATE*–estimator is biased when covariates are stochastic. Therefore, the well-known formulas for ordinary least-squares regressions cannot be used to derive valid statistical inference for generalized analysis of covariance. Our derivation explains underestimated standard errors observed for the test of the average total effect based on the general linear model (see Flory, 2004). Estimating generalized analysis of covariance based on mean-centered covariates simplifies the computation of the average total effect but it will not change the underestimation of the corresponding standard error estimators.

Probably because the conditional and the unconditional approaches of regression share a common nomenclature, the distinction of stochastic versus fixed regressors is not very common in the regression literature. For the testing of average total effects based on covariate-treatment regressions as considered in this thesis, the incorporation of the stochastic nature of the covariates is of importance.

Note that we have made no assumption about the joint distribution of regressors and the outcome variable for deriving the condition under which the variance of the usual ordinary least-squares estimator for the average total effect is underestimated. Nevertheless, as argued by Maddala (1992), we need to make assumptions about the joint distribution of Y, Z and X in order to derive valid standard errors. Instead of deriving a corrected standard error estimator based on ordinary least-squares regressions under a specific distributional assumption, we will utilize structural equation modeling for the implementation of a correct test statistic for the (unconditional inference about the) average total effect in the next section. The reason is basically that within the framework of structural equation modeling a multivariate distributional assumption is routinely made and we therefore can rest on the easily available methodology for a gener-

alized analysis of covariance.[20] A similar strategy was suggested by Allison (1995) for testing whether the change in a regression coefficient is statistically significant when a regressor is added to a regression model. The author derived the conditions under which the stochasticity of the regressors is important for testing these hypotheses and developed a test statistic based on the assumption of a multivariate joint distribution. Finally, Allison (1995) suggested a likelihood ratio test with the help of nonlinear constraint structural equation models.

3.3 Structural Equation Modeling

Structural equation modeling has already been successfully used as an alternative to classical multivariate methods, for example, to generalize the analysis of group differences in factor means (MANOVA) [Sörbom, 1974; Bray & Maxwell, 1985; Bagozzi & Yi, 1989; Cole, Maxwell, Arvey, & Salas, 1993]. Following this strategy we shall now discuss the implementation of generalized analysis of covariance as structural equation model with nonlinear constraints in this section.

3.3.1 Introduction

This section is organized as follows: We shall start with a brief description of the statistical model underlying (multi-group) structural equation mosels. Additionally, we shall describe different strategies to test hypotheses about nonlinear functions of parameters within the framework of structural equation modeling. This will be followed by a detailed development of different implementations of generalized analysis of covariance as structural equation model with nonlinear constraints.

Statistical Model The statistical model underlying structural equation models in the case of continuous observed variables is given by two separate equations for each group $g = 1, ..., G$ (see L. K. Muthén & Muthén, 1998 - 2007, for the `Mplus`–framework, and also Jöreskog & Sörbom, 1996 - 2001, for an equivalent notation in the well known `LISREL` notation). The measurement model regresses the vector of observed variables $\boldsymbol{y}_i^{(g)}$ on a vector of latent variables $\boldsymbol{\eta}_i^{(g)}$ [upper part in Equation (3.36)], and the structural model regresses the latent variable on each other (lower part):

$$\begin{aligned} \boldsymbol{y}_i^{(g)} &= \boldsymbol{\nu}^{(g)} + \boldsymbol{\lambda}^{(g)} + \boldsymbol{\eta}_i^{(g)} + \boldsymbol{\epsilon}^{(g)} \quad \text{(measurement model)} \\ \boldsymbol{\eta}_i^{(g)} &= \boldsymbol{\alpha}^{(g)} + \boldsymbol{B}^{(g)}\boldsymbol{\eta}_i^{(g)} + \boldsymbol{\zeta}_i^{(g)} \quad \text{(structural model)}. \end{aligned} \tag{3.36}$$

[20] Note that structural equation modeling itself is sometimes described as *causal modeling* (see Bentler, 1980) because of possible confirmatory and hypothesis-testing applications of this methodology (Jöreskog, 1969). A comprehensive summary of this research tradition can be found for instance in Mulaik (2009).

The overall implied variance-covariance matrix and the overall implied mean vector are composed of the group-specific contributions. Based on Equation (3.36), the group-specific implied mean structure $\boldsymbol{\mu}_y^{(g)}$ of the observed manifest variables is expressed as:

$$\boldsymbol{\mu}^{(g)} \quad = \quad \boldsymbol{\nu}^{(\boldsymbol{g})} + \boldsymbol{\Lambda}^{(\boldsymbol{g})} \left(\boldsymbol{I} - \boldsymbol{B}^{(\boldsymbol{g})}\right)^{-\mathbf{1}} \boldsymbol{\alpha}^{(\boldsymbol{g})}. \tag{3.37}$$

Furthermore, the model implied covariance structure $\Sigma_y^{(g)}$ can be written as

$$\boldsymbol{\Sigma}_y^{(g)} \quad = \quad \boldsymbol{\Lambda}^{(\boldsymbol{g})} \left(\boldsymbol{I} - \boldsymbol{B}^{(\boldsymbol{g})}\right)^{-\mathbf{1}} \boldsymbol{\Psi}^{(\boldsymbol{g})} \boldsymbol{\Lambda}^{(\boldsymbol{g})} \left(\boldsymbol{I} - \boldsymbol{B}^{(\boldsymbol{g})}\right)^{'-\mathbf{1}} \boldsymbol{\Lambda}^{'(\boldsymbol{g})} + \boldsymbol{\Theta}^{(\boldsymbol{g})}, \tag{3.38}$$

with $\boldsymbol{\Psi}^{(g)}$ as the matrix of covariance between the residuals of the structural model ($\boldsymbol{\zeta}_i^{(g)}$) as well as variances and covariances between the residuals of the measurement model $\boldsymbol{\epsilon}_i^{(g)}$ in $\boldsymbol{\Theta}^{(g)}$.

In the literature on structural equation modeling, the typical notation combines all parameters of the model into a vector $\boldsymbol{\theta}$ (see, e. g., Bollen, 1989). Estimates ($\widehat{\boldsymbol{\theta}}$) for the parameter in $\boldsymbol{\theta}$ are obtained by minimizing a fitting function. The multi-group *fitting function* for the ML estimator for continuous observed variables is given, e. g., by B. Muthén (1998-2004) as

$$F_{\text{ML}} = 1/2 \sum_{g=1}^{G} \left(n_g \left[ln \left| \boldsymbol{\Sigma}_y^{(g)}(\boldsymbol{\theta}) \right| + tr \left(\boldsymbol{\Sigma}_y^{(g)\,-1}(\boldsymbol{\theta}) \mathbf{T}_g(\boldsymbol{\theta}) \right) - ln \left| \mathbf{S}_g \right| \right] - (p+q) \right) / n, \tag{3.39}$$

with $p+q$ as the number of observed variables, n as the total sample size, G as the number of groups, and n_g as the sample size in group g. The matrix $\mathbf{T}_g(\boldsymbol{\theta})$ is a group-specific composite of the observed variance-covariance matrix $\mathbf{S}_g$ and the squared derivation of the observed $[\overline{\boldsymbol{\nu}}^{(g)}]$ and the implied $[\boldsymbol{\mu}^{(g)}(\boldsymbol{\theta})]$ mean structure:

$$\mathbf{T}_g(\boldsymbol{\theta}) = \mathbf{S}_g + \left(\overline{\boldsymbol{\nu}}^{(g)} - \boldsymbol{\mu}^{(g)}(\boldsymbol{\theta}) \right) \left(\overline{\boldsymbol{\nu}}^{(g)} - \boldsymbol{\mu}^{(g)}(\boldsymbol{\theta}) \right)^T. \tag{3.40}$$

Provided that the model is identified, parameter estimates obtained by minimizing Equation (3.39) yield a minimal difference between the implied variance-covariance matrix $\boldsymbol{\Sigma}_y^{(g)}(\boldsymbol{\theta})$ of the manifest variables $\boldsymbol{y}_i^{(g)}$ and the observed variance-covariance matrix $\mathbf{S}_g$, as well as the difference between the implied mean vector $\boldsymbol{\mu}^{(g)}(\boldsymbol{\theta})$ and the empirical mean vector $\overline{\boldsymbol{\nu}}^{(g)}$ (see, e. g., Steyer, Wolf, Funke, & Partchev, 2009).

Distributional Assumptions Multivariate normality of the observed variables is a typical assumption for the parameter estimation within structural equation modeling. For example, the common maximum likelihood estimator F_{ML} [see Equation (3.39)] is developed under the assumption of multivariate normally distributed observed variables (that is, under the assumption that the sample covariance matrix $\mathbf{S}_g$ follows a *Wishart Distribution*, see Bollen, 1989, p. 134), i. e., a multivariate normality conditional on the

grouping variable (X) is assumed for the multi-group estimator. When these distributional assumptions are met, standard errors are asymptotically unbiased estimators of the variability of the estimated parameters (Bollen, 1989). Asymptotically, this is also valid for the standard errors of functions of (estimated) parameters (see below).

Specification Error Biases due to model misspecifications in structural equation models have been discussed in the literature for a long time because *a misspecified model can result in biased estimation* (Gallin, 1983). The bias due to model misspecification has been termed *specification error* in the literature (see, for example, Wold, 1956; Farley & Reddy, 1987). For structural equation models, Yuan, Marshall, and Bentler (2003, p. 241) note that *"model misspecification may have a systematic effect on parameters, causing biases in their estimates"*, and for instance, Kaplan (1989) studied the effect of specification error on the estimated standard errors and reported the finding that misspecification in one parameter can affect the estimated standard errors of this parameter and also the estimated standard errors for other parameters of the model (see also Curran, West, & Finch, 1996). Specification error in structural equation models is often discussed as a property of the estimation method (see, for example, Olsson, Foss, Troye, & Howell, 2000, for a comparison of the performance of maximum likelihood, generalized least-squares, and weighted least-squares under misspecification).

Directly related to the implied variance structure, Bollen (1996) discussed an alternative estimation method for structural equation models with heteroscedastic error variances of an unknown form (a limited information approach, called *two-stage least-squares estimator*, 2SLS; see also Bollen & Paxton, 1998). As described in section 3.1.2, different sources for heterogeneity of residual variances and heteroskedasticity can be distinguished for the specification of the covariate-treatment regression. For the implementation of generalized analysis of covariance under the assumption of Z-conditional unbiasedness and with an appropriate functional form assumption, we will study how to take the X–related heterogeneity of residual variances for the specification of the model into account. Furthermore, as B. O. Muthén and Asparouhov (2003) argue, the ability of structural equation models to use full information maximum-likelihood estimation instead of the limited information approach increases efficiency and gives power advantages for subsequent test statistics. We will therefore not study the performance of 2SLS within this thesis (for different latent variable models, comparisons can be found in the literature, for example, in Kaplan, 1988).

Summary Structural equation modeling seems to be a promising framework for estimating and testing average total effects based on a generalized analysis of covariance (see also Steyer & Partchev, 2008) for quasi-experimental designs because of the following features: Technically, as highlighted in section 3.1.3,

the common assumption of a joint (multivariate) distribution is necessary to incorporate the stochastic nature of the regressors (X and Z) in the moderated (latent) regressions appropriately.

Moreover, the theoretically motivated heterogeneity of residual variances can easily be accounted for, for instance, because of the capabilities of multi-group modeling based on group-specific variance-covariance matrices (Sörbom, 1978, see also Arbuckle, 2006). Even if no latent variables are involved, the framework of structural equation modeling provides an interesting alternative to the general linear model, e. g., the flexibility of standard software to test specialized and combined hypotheses about different model parameters and the availability of modern techniques for handling missing values (in particular due to the *full information maximum likelihood* method, see, for example, Arbuckle, 1996; Enders, 2001a).

Finally, the ability to handle latent variables is probably the most distinct feature of generalized analysis of covariance implemented as structural equation model. For covariate-treatment regression without interaction terms, the simple multi-group model suggested by Sörbom (1978) is capable of accounting for measurement error of the covariates. As we have already discussed in subsection 2.5.5, although none of the research questions studied in this thesis deal with the effect of measurement error, the structural equation models which will be developed and studied here can be applied as adjustment methods with latent covariates (see Steyer & Partchev, 2008, and Steyer et al., in press).

Overall, we will make use of the available methodology for structural equation models to weaken assumptions of the general linear model. Fur multi-group structural equation models this is obvious with respect to the heterogeneity of residual variances, but it needs to be discussed with respect to the fixed-X assumption (because multivariate normality is assumed conditional on X, see subsection 3.3.1).

3.3.2 Testing Nonlinear Constraints in Structural Equation Models

We are now going to summarize different strategies to test hypotheses (for instance, about average total effects) within the framework of structural equation modeling. This short survey is included here to show the similarity between the Wald–test used in Steyer and Partchev (2008) and the test statistic applied by Flory (2008). The basic statistical model for the structural equation modeling approach was already introduced in section 3.1.3 (see Bollen, 1989, for details). For the discussion of different test statistics for hypotheses about the average total effect, we assume that the parameters of the covariate-treatment regression are estimated with a maximum likelihood (ML) based estimation method by minimizing, e. g., Equation (3.39), or by the more general full information maximum likelihood equivalent (Arbuckle, 1996).

Within the framework of structural equation modeling different strategies for testing hypotheses about constraints have been developed and adapted from general statistical theory. The following different approaches to test hypotheses about the average total effect are described in the next subsection: The *likeli-*

hood ratio test, the *Wald–test*, the *Lagrange Multiplier test*, and the test statistic based on the standard error of the estimated average total effect (z–test).

Nonlinear Constraints Statistical inference about the average total effect based on a covariate-treatment regression with covariate-treatment interaction can be drawn with the help of nonlinear constraints within the framework of structural equation modeling. Providing an appropriate model specification for the covariate-treatment regression, we can express the hypothesis $H_0 : ATE_{10} = 0$ as a *nonlinear constraint*, i. e., as a function of the estimated model parameters:

$$H_0 : c(\widehat{\boldsymbol{\theta}}) = 0. \tag{3.41}$$

Technically, different constraints can be specified for a *constrained estimation* of structural equation models, i. e., for the minimization of Equation (3.39) with additional restrictions (see for a summary of the different constraints, e. g., Kline, 2005). To differentiate the general linear hypothesis, which is assumed to be *linear* in its parameters (see, subsection 3.2.3), we use the term *nonlinear constraint* to emphasize that $c(\widehat{\boldsymbol{\theta}})$ might be a nonlinear function of (estimated) model parameters. However, as explained in the following paragraph, we do not generally claim that a *constrained estimation* is performed with respect to the nonlinear constraint in Equation (3.41). Instead, when referring to the average total effect we use the term nonlinear constraint as a synonym for the nonlinear function of estimated model parameters of the covariate-treatment regression (which yields an estimator of the average total effect).

Likelihood Ratio Test Most computer packages for structural equation modeling are capable of estimating constrained structural equation models (for example `LISREL`, Jöreskog & Sörbom, 1996 - 2001, `EQS`, Bentler, 1995, and `Mplus`, L. K. Muthén & Muthén, 1998 - 2007). With this option, the parameters of the structural equation model are estimated by simultaneously minimizing Equation (3.39) and satisfying Equation (3.41) [see, for example, Tang & Bentler, 1998]. Hence, a test statistic based on the value of the likelihood function for the restricted estimation of $\widehat{\theta}$ (L_R) and the value of the likelihood function for the estimation of $\widehat{\theta}$ without the restriction (L_U) can be constructed. Comparing the two values of the likelihood function allows one to test hypotheses about the average total effect. This strategy is known as the likelihood ratio test (sometimes referred to as the χ^2-difference test). Given that the null hypothesis is true, the following test statistic

$$\text{LR} = -2ln\left(\frac{L_R}{L_U}\right) \tag{3.42}$$

is χ^2–distributed for sufficiently large samples, with degrees of freedom equal to the number of restrictions imposed by $c(\widehat{\boldsymbol{\theta}})$ [see, for the general properties of the likelihood ratio test, Greene, 2007, and Bollen, 1989,

for the likelihood ratio test for latent variable models]. The underlying idea is that if the restrictions imposed by the nonlinear constraint are valid, they should not lead to a large reduction in the value of the log-likelihood function. The ratio $\frac{L_R}{L_U}$ must be between zero and one because both likelihoods are positive and the unrestricted optimum is always superior to the restricted one.

Wald Test A practical shortcoming of the likelihood ratio test described in the last paragraph is that the estimation of a restricted model is required to obtain L_R in addition to the estimation of the unrestricted model for L_U. The constrained model is misspecified provided that the null hypothesis is true. For that reason, the risk of non-convergence is high and likelihood ratio tests are inconvenient to some degree.

Hypotheses about the average total effect can be tested alternatively based on the unconstrained model only. The underlying idea is generally known as the *Wald–test* (Wald, 1943). By translating the hypothesis in Equation (3.41) in to the form $c(\widehat{\boldsymbol{\theta}}) = \mathbf{q}$, that is with $\mathbf{q} = 0$, the following test statistic can be derived

$$\text{Wald} = \left[c(\hat{\boldsymbol{\theta}}) - \mathbf{q}\right] \left[acov\left(c(\hat{\boldsymbol{\theta}}) - \mathbf{q}\right)\right]^{-1} \left[c(\hat{\boldsymbol{\theta}}) - \mathbf{q}\right], \tag{3.43}$$

where $acov\left(c(\widehat{\boldsymbol{\theta}}) - \mathbf{q}\right)$, is the asymptotic variance-covariance matrix of the constraint obtained from the estimation of L_U. Under the null hypothesis and for large samples, the test statistic in Equation (3.43) is χ^2–distributed with degrees of freedom equal to the number of restrictions imposed by $c(\widehat{\boldsymbol{\theta}}) = \mathbf{q}$. The Wald–test is asymptotically equivalent to the likelihood ratio test (DasGupta, 2008).

Lagrange Multiplier Test In contrast to the Wald–test (which is based on the unrestricted L_U), the restricted L_R is utilized for the *Lagrange multiplier test*. In textbooks about structural equation modeling, the Lagrange multipliers are well known as *modification indices* (see, for example, Kaplan, 2000). As described in detail by Greene (2007), two terms are necessary to compute the Lagrange multiplier test:

$$\text{LM} = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right)^T \left[\mathbf{I}(\hat{\theta}_R)\right]^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right). \tag{3.44}$$

The matrix $\mathbf{I}\left(\widehat{\boldsymbol{\theta}}_R\right)$ denotes the information matrix, which is available for the maximum likelihood estimation within the framework of structural equation modeling [see Equation (3.47)]. The term $\frac{\partial \ln L(\hat{\theta}_{\mathbf{R}})}{\partial \hat{\theta}_{\mathbf{R}}} = -\widehat{\mathbf{C}}^T \hat{\lambda}$ is zero if the constraint is valid, where $\hat{\lambda}$ is the vector of estimated Lagrange multipliers (modification indices as computed by conventional program packages for the analysis of structural equation models), and $\widehat{\mathbf{C}}$ is the matrix of partial derivatives of the constraint with respect to the model parameters [see Equation (3.49)].

The available program packages for the estimation of structural equation models do not give estimates of the Lagrange multipliers for every model developed in section 3.3 at this time. Furthermore, as mentioned above, the estimation of the restricted model is prone to non-convergence problems.

Test based on the Standard Error A fourth option to test the hypothesis of an average total effect different from zero is the so-called z–test based on the (estimated) standard error of the nonlinear constraint. For this test, the estimated average total effect, as a function of model parameters, is divided by its estimated standard error. The ratio of a parameter estimate and its standard error is asymptotically normally distributed (see, e. g., Wasserman, 2004). This property can be used to establish significance tests, or to construct a confidence interval for model parameters of structural equation models (see, e. g., Bollen, 1989):

$$z = \frac{|\widehat{ATE}_{10} - ATE_{10}|}{\text{S.E.}(\widehat{ATE}_{10})}. \tag{3.45}$$

For the simple two group case of generalized analysis of covariance considered in this thesis, the nonlinear constraint yields a single restriction for the maximum likelihood estimation and hence the z–test is equivalent to the Wald–test described in the last paragraph. This equivalence can easily be verified because $c(\boldsymbol{\theta}) - q = 0$ can be substituted by $\widehat{ATE}_{10} - ATE_{10}$ for the hypothesis $ATE_{10} = 0$:

$$\begin{aligned} \text{Wald} \quad &= \left[c(\hat{\boldsymbol{\theta}}) - q\right] \left[avar\left(c(\hat{\boldsymbol{\theta}}) - q\right)\right]^{-1} \left[c(\hat{\boldsymbol{\theta}}) - q\right] \\ &= \left[\widehat{ATE}_{10} - ATE_{10}\right] \left[acov\left(\widehat{ATE}_{10} - ATE_{10}\right)\right]^{-1} \left[(\widehat{ATE}_{10} - ATE_{10})\right] \\ &= \frac{\left(\widehat{ATE}_{10} - ATE_{10}\right)^2}{acov\left[\widehat{ATE}_{10}\right]} = \frac{\left(\widehat{ATE}_{10} ATE_{10}\right)^2}{\text{S.E.}(\widehat{ATE}_{10})^2} = z^2. \end{aligned} \tag{3.46}$$

As given by Greene (2007), the test statistic of the Wald–test follows a χ^2–distribution with one degree of freedom, which is the square of the standard normal distribution of z in Equation (3.45).

For some of the structural equation models developed in this thesis, the standard error of the nonlinear constraint, i. e., the standard error of the *ATE*–estimator, is not easily available without additional assumptions. Accordingly, for quasi-experimental designs where the treatment variable X is a stochastic regressor, the derivation of an approximated standard error for the estimated average total effect was suggested by Nagengast (2006). In order to discuss the underlying assumption of this approach, we shall provide a review of the (multivariate) δ-method in the following paragraph.

(Multivariate) δ-method The δ-method is a very useful tool to derive the variance of a function of a random variable (Rao, 1973, p. 388, see also Oehlert, 1992, and Raykov & Marcoulides, 2004). This statistical

tool can be applied generally for a random variable whose distribution depends on a real-valued parameter and for any function of the random variable which can be differentiated with respect to the parameter.

In its multivariate extension, the method involves two parts: the (asymptotic) variances and covariances of the incorporated random variables and the partial derivatives of the functions of the random variables with respect to the parameters (see, e. g., D. P. MacKinnon, 2008, p. 91 ff., for a non-technical description and Wasserman, 2004, for a comprehensive discussion). The variance of smooth functions of model parameters is approximated with the multivariate δ-method based on a first-order Taylor expansion (Bishop, Fienberg, & Holland, 1975, p. 487).

The asymptotic variance-covariance matrix of the estimated (unconstrained) parameters, $acov(\hat{\boldsymbol{\theta}})$, is the starting point for calculating a standard error for the estimated average total effect. Bollen (1989, p. 109 and appendix 4B therein) provided the formula for the asymptotic variance-covariance matrix for structural equation models estimated by the maximum likelihood fitting function [see Equation (3.39)] as

$$acov(\hat{\boldsymbol{\theta}}) = \left(\frac{2}{N-1}\right)\left\{E\left[\frac{\partial^2 F_{ML}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]\right\}^{-1}. \tag{3.47}$$

For the smooth function $f(\hat{\boldsymbol{\theta}}) = c(\hat{\boldsymbol{\theta}}) = \widehat{ATE}_{10}$ of model parameters, the asymptotic variance-covariance matrix is estimated from $\widehat{acov}\left(\hat{\boldsymbol{\theta}}\right)$ as

$$\widehat{acov}\left(c(\hat{\boldsymbol{\theta}})\right) \quad = \hat{\mathbf{C}}\,\widehat{acov}\left(\hat{\boldsymbol{\theta}}\right)\,\hat{\mathbf{C}}^T \tag{3.48}$$

by pre- and post-multiplying with

$$\hat{\mathbf{C}} = \frac{\partial c(\hat{\boldsymbol{\theta}})}{\partial\hat{\boldsymbol{\theta}}^T} \tag{3.49}$$

as the $J \times K$ matrix of partial derivatives of the constraint with respect to the K elements of the parameter vector θ (called the *Jacobian*), where K is the total number of parameters in the structural equation model and J is the number of groups (see, e. g., Raykov & Marcoulides, 2004, p. 628).

The standard error for the function of parameters (in our application the standard error of the *ATE*–estimator) is obtained as the square root of the asymptotic variance

$$\widehat{\text{S.E.}}(\widehat{ATE}_{10}) = \sqrt{\widehat{acov}\left[\widehat{ATE}_{10}\right]} = \sqrt{\widehat{acov}\left[c(\hat{\boldsymbol{\theta}})\right]}. \tag{3.50}$$

The δ-method reviewed here can be applied if an estimate of the variance-covariance matrix of the parameter estimates is available for all parameters involved in the constraint.

Summary The Lagrange multiplier test and the likelihood ratio test are based either on the restricted, or the restricted and the unrestricted model. As argued, this might be disadvantageous as the restricted model should by theory be misspecified under the null hypothesis. The Wald–test can be applied to test multiple or combined hypotheses based on the estimated parameters and their asymptotic variance-covariance matrix. Therefore, we will focus on the Wald–test as the most flexible tool, which is equivalent to the test based on the standard error for all structural equation models studied in this thesis. This means that the statistics considered for the average total effect are at their core based on the asymptotic variance-covariance matrix of the parameter estimates, $acov(\widehat{\boldsymbol{\theta}})$, and the value of the function $c(\widehat{\boldsymbol{\theta}})$.

3.3.3 Specification of Multi-Group Structural Equation Models

The following two sections present five different structural equation models for the implementation of generalized analysis of covariance. We will start with the multi-group implementation of the analysis of covariance as a structural equation model, the *simple multi-group model*, which goes back to Sörbom (1978). With respect to the implementation of the introduced hypothesis $H_0 : ATE_{10} = c(\widehat{\boldsymbol{\theta}}) = 0$ as a nonlinear constraint (see subsection 3.3.1) we will demonstrate that the straight generalization of Sörbom's model for covariate-treatment regressions with interaction terms is unfeasible due to the stochasticity of the treatment variable X. We will then describe a slightly different implementation which takes the randomness of X into account (labeled as *elaborated multi-group model*), and summarize the *approximated multi-group model*, which has already been implemented in the software *EffectLite* (Steyer & Partchev, 2008).

In section 3.3.4 we will summarize the *simple single group model* which was studied by Flory (2008), and present our extension, the *elaborated single group model*, as an implementation of generalized analysis of covariance that might be more appropriate when heterogeneity of residual variances is expected.

3.3.3.1 Simple Multi-Group Model (with Fixed Group Size)

The basic idea of specifying the analysis of covariance as a structural equation model for simultaneously analyzing data from different groups was introduced in a highly influential paper by Sörbom (1978). The underlying statistical property allowing this kind of multi-group analysis is the additivity of the log-likelihood functions [i. e., the sum over group-specific log-likelihood functions in Equation (3.39), see Bollen, 1989]. Note that the additivity of the log-likelihood functions also serves as the starting point for the derivation of the full-information maximum likelihood method (FIML, see e. g., Davey & Savla, 2010; Enders, 2001b; Wolf, 2006; Kröhne & Wolf, 2002). Nevertheless, the multi-group analysis of covariance as suggested by Sörbom (1978) needs to be extended for average total effects when covariate-treatment interactions are present, as we will show in the next paragraph.

Model Specification For the presentation of generalized model, the simple linear parameterized covariate-treatment regression [see Equation (1.29) on page 17] serves again as the starting point. Instead of an application of ordinary least-squares regression, we shall now discuss the estimation of this covariate-treatment regression within the framework of structural equation modeling, implemented as a simple multi-group model. Group-specific variance-covariance matrices are used to estimate the parameters of the regressions $E_{X=0}(Y|Z) = \beta_{00} + \beta_{01} \cdot Z$ and $E_{X=1}(Y|Z) = \beta_{10} + \beta_{11} \cdot Z$, i. e., the treatment group-specific covariate regressions. A path diagram of this model for the case of only one manifest covariate Z is given in Figure 3.1 and the corresponding `Mplus`–syntax is shown in Listing 3.1 (see also Steyer et al., in press, and Steyer & Partchev, 2008).

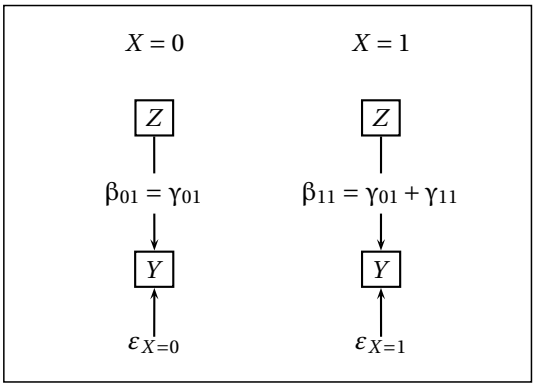


Figure 3.1: Path diagram of the simple multi-group model with fixed group size

```
DATA:     FILE = example.dat;
VARIABLE: NAMES ARE Y X Z;
          USEVARIABLES ARE Z Y;
          GROUPING = X (1=T 0=C);
ANALYSIS: TYPE IS MEANSTRUCTURE;
MODEL:    Y ON Z;
MODEL C:  Y ON Z(beta01);         ! E(Y|Z,X=0)
          [Y](beta00);            ! (Control)
          [Z](EzX0);
          Y; Z;
MODEL T:  Y ON Z(beta11);         ! E(Y|Z,X=1)
          [Y](beta10);            ! (Treatment)
          [Z](EzX1);
          Y; Z;

MODEL CONSTRAINT:   new(ATE10,Ez);
                    Ez = (EzX1*0.50)+(EzX0*0.50);    ! P(X=1) := 0.50
                    ATE10 = beta10 - beta00
                           + (beta11 - beta01) * Ez;

MODEL TEST:         0 = ATE10;
```

Listing 3.1: `Mplus`–syntax for the simple multi-group model

The regression coefficients for the group-specific regressions of the outcome variable Y on the covariates Z and the conditional (i. e., group-specific) expectations of the covariate $E(Z|X = j)$ are estimated as model parameters by normal theory maximum-likelihood [see Equation (3.39) and, e. g., Bollen, 1989, for a general overview, as well as B. Muthén, 1998-2004, Appendix 4 for technical details] under the distributional assumptions mentioned above.

According to this multi-group parameterization of the covariate-treatment regression, the average total effect obtained as the expectation of the effect function is

$$\begin{aligned} ATE_{10} &= E\big(g_1(Z)\big) \\ &= E\big((\beta_{10}-\beta_{00})+(\beta_{11}-\beta_{01})\cdot Z\big) \\ &= \big(\beta_{10}-\beta_{00}\big)+\big(\beta_{11}-\beta_{01}\big)\cdot E(Z) \\ &= \big(\beta_{10}-\beta_{00}\big)+\big(\beta_{11}-\beta_{01}\big)\cdot\big(E_{X=0}(Z)P(X=0)+E_{X=1}(Z)P(X=1)\big), \end{aligned} \tag{3.51}$$

provided that the covariate-treatment regression is Z-conditional unbiased and that the functional form assumption for $E_{X=j}(Y|Z)$ holds for each value j of X. Note that the unconditional expectation $E(Z)$ is replaced with the conditional expectations, weighted by the treatment probabilities, i. e.,

$$E(Z)=\sum_{j=0}^{J-1} E_{X=j}(Z)\cdot P(X=j). \tag{3.52}$$

For generalized analysis of covariance with covariate-treatment interaction implemented as the simple multi-group model in `Mplus`, the average total effect estimator is obtained as the (nonlinear) function of the estimated regression coefficients and the unconditional expectation of the covariate Z (see lines 16-19 in listing 3.1) and tested with the Wald–test (see line 21). This test is, as summarized in the last section, identical to the test based on the standard error of the estimate of the average total effect.

Stochasticity of Z and X For the analysis of the consequences of stochastic covariates on inferences about the average total effect in subsection 3.2.5, we assumed access to $\hat{\mu}_Z$ in Equation (3.27) as the unconditional mean of the covariate Z, either as a model parameter or as a known number. Finally, we found out that the variance of the *ATE*–estimator is underestimated if $\hat{\mu}_Z$ is an estimated quantity with associated variance [and if the regression slopes are different between the groups, see Equation (3.32)]. We shall now proceed and discuss the consequences of stochastic treatment variables for structural equation modeling, that means we shall consider the properties of the mean of the group indicator variable $\hat{\mu}_{I_{X=j}}$, which is assumed to be an unbiased estimate for the treatment probability, i. e., $E(\hat{\mu}_{I_{X=j}}) = P(X = j)$ [the *group size*].[21]

[21] The group indicator variable $I_{X=j}$ is defined as $I_{X=j} \equiv \begin{cases} 1, & \text{if } X=j \\ 0, & \text{otherwise} \end{cases}$ for $j = 0,...,J-1$.

The parameters of the model can be estimated based on the maximum likelihood estimator mentioned in section 3.3.1 by minimizing the fitting function in Equation (3.39), taking into account the joint distribution of Y and Z, conditional on X. As we derived from the literature on ordinary least-squares regression, the incorporation of this joint distribution is necessary to draw valid unconditional inference about the average total effect.

Nevertheless, the simple multi-group model does not incorporate the stochastic nature of the regressor X when covariate-treatment interactions are incorporated.[22] To illustrate this drawback, we will describe the translation of the covariate-treatment regression into the statistical model in more detail, based on the presentation of the multivariate δ-method given in section 3.3.1 on page 79. Technically, the covariate-treatment regression with interaction term [see Equation (1.29) on page 17] fits into the general model given in Equation (3.37) and Equation (3.38) as group-specific regressions $E_{X=0}(Y|Z) = \beta_{00} + \beta_{01} \cdot Z$ and $E_{X=1}(Y|Z) = \beta_{10} + \beta_{11} \cdot Z$. No additional equality constraints for the slopes are specified. Accordingly, the regression coefficients β_{00} and β_{10} are (freely) estimated as $\hat{\alpha}_1^{(1)}$ and $\hat{\alpha}_1^{(2)}$; β_{01} and β_{11} are estimated as $\hat{b}_1^{(1)}$ and $\hat{b}_1^{(2)}$, and the conditional expectations of the covariate given treatment $E_{X=0}(Z)$ and $E_{X=1}(Z)$ are estimated as $\hat{\alpha}_2^{(1)}$ and $\hat{\alpha}_2^{(2)}$.[23]

The formulation of an estimator based on the expression for the average total effect in Equation (3.51) reveals that the statistical model underlying the simple multi-group model does not account for the stochastic group membership. The estimator of the average total effect for a covariate-treatment regression without interaction [see Equation (3.53)] is obtained as $\widehat{ATE}_{10} = \hat{\alpha}_1^{(2)} - \hat{\alpha}_1^{(1)}$ without difficulties. When the regression slopes are not parallel (i. e., $\beta_{11} \neq \beta_{01}$), the estimator of the average total effect [see Equation (3.51)] must include either an estimate for the unconditional expectation of the covariate $E(Z)$ or the estimated conditional (i. e., group-specific) expectations $E_{X=j}(Z)$ as well as the estimated group sizes $P(X = j)$. Unfortunately, only the conditional expectations of the covariate are included and estimated in the model. Therefore, we cannot express the average total effect as a smooth function of estimated model parameters. Accordingly, we can neither compute a confidence interval for the estimate nor conduct a test for the hypothesis $H_0 : ATE = 0$ based on the standard error of the estimator. Consequently, it is also impossible to perform a Wald–test of the hypothesis based on the asymptotic covariance matrix of the constraint, or

[22] Note that for the model suggested by Sörbom (1978), i. e., for a covariate-treatment regression without interaction term (with $\beta_{11} = \beta_{01}$), the unconditional expectation $E(Z)$ is not necessary to formulate the constraint:

$$\begin{aligned} ATE_{10} &= (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01}) \cdot E(Z) \\ &= \beta_{10} - \beta_{00}. \end{aligned} \tag{3.53}$$

[23] The following elements of θ are estimated for the simple multi-group model based on the `Mplus`–syntax of listing 3.1:

$$\theta = \begin{pmatrix} \beta_{00} & E_{X=0}(Z) & \beta_{01} & Var(\varepsilon_{X=0}) & Var_{X=0}(Z) & \beta_{10} & E_{X=1}(Z) & \beta_{11} & Var(\varepsilon_{X=1}) & Var_{X=1}(Z) \end{pmatrix}^T. \tag{3.54}$$

to fit a restricted model where the average total effect is constrained to zero, as would be necessary for a likelihood ratio test.

In other words, the estimator of the average total effect cannot be formulated as a function of estimated model parameters of the standard `Mplus`–model for the general case of a covariate-treatment regression with interaction term.[24] This is due to the simple fact that although the $\widehat{\beta}$ coefficients are estimated as part of the parameters $\widehat{\boldsymbol{\theta}}$, the unconditional expectation of the covariate $E(Z)$ is not included and must be replaced by the conditional expectations $E_{X=j}(Z)$ weighted with the treatment probabilities. If these treatment probabilities $P(X=j)$ are assumed to be known, the estimator of the average total effect can be expressed as a function of model parameters and test statistics can be derived:

$$\begin{aligned}
\widehat{ATE}_{10} &= \left(\hat{\beta}_{10}-\hat{\beta}_{00}\right)+\left(\hat{\beta}_{11}-\hat{\beta}_{01}\right)\cdot\widehat{\mu}_Z \\
&= \left(\hat{\beta}_{10}-\hat{\beta}_{00}\right)+\left(\hat{\beta}_{11}-\hat{\beta}_{01}\right)\cdot\left(\widehat{\mu}_{Z|X=0}\cdot P(X=0)+\widehat{\mu}_{Z|X=1}\cdot P(X=1)\right) \\
&= \left(\hat{\alpha}_1^{(2)}-\hat{\alpha}_1^{(1)}\right)+\left(\hat{b}_1^{(2)}-\hat{b}_1^{(1)}\right)\cdot\left(\hat{\alpha}_2^{(1)}\cdot P(X=0)+\hat{\alpha}_2^{(2)}\cdot P(X=1)\right).
\end{aligned} \tag{3.55}$$

The validity of the statistical inference for the estimator in Equation (3.55) under the assumption that the treatment probability $P(X=1)$, and therefore also $P(X=0)=1-P(X=1)$, are fixed known numbers, depends on the underlying design of the study. For observational studies considered in this thesis, where X is not completely controlled by design, the δ-method cannot be applied to derive the variance of the estimator in Equation (3.55) because comparable to the stochasticity of Z described in section 3.2.5, the uncertainty for the estimation of $P(X=1)$ is not appropriately accounted for.

According to Flory (2008), we distinguish between two versions of the simple multi-group model: The *simple multi-group model (sample)* is obtained by substituting the sample estimates of the group sizes (the observed frequencies, ignoring stochasticity) in Equation (3.52). In the Monte-Carlo simulation to be described in chapter 4, we present our study of the same model based on the true population values of the group size used for generating the data. We will label the resulting approach as *simple multi-group model (population)*. Subject to this choice of *sample* or *population*, the corresponding fixed numbers (either the sample estimates of the treatment probability, or the true assignment probability used for generating the data) are inserted in the constraint in line 16-20 of listing 3.1.

Neither alternative simple multi-group models are satisfying: The true group sizes are rarely known in quasi-experimental designs, i. e., this approach is often not feasible. The estimated group size is available for empirical applications, but the treatment of this estimate is not appropriate in the simple multi-group model (sample). We will illustrate this inappropriateness of the simple multi-group model with the help of

[24] Since the same is true for the `LISREL` model as well, we omit the details here.

the Monte Carlo study, for which the treatment variable and the covariate will be generated as stochastic regressors.

Heterogeneity of Residual Variances It is important to note that the residual variances $Var(\varepsilon_{X=0})$ and $Var(\varepsilon_{X=1})$ need not necessarily be equal for the specification of the multi-group models considered in this thesis. Technically, we do not specify equality constraints for the corresponding diagonal elements of $\boldsymbol{\Psi}^{(g)}$ [see listing 3.1, and Equation (3.38)]. As a consequence, these two variances are both part of the parameter vector $\boldsymbol{\theta}$ (see footnote 23) and are estimated by minimizing the multi-group fitting function [see Equation (3.39)]. If we considered the implied variance-covariance structure of the multi-group models, the freely estimated residual variances would add to the implied group-specific variances of the outcome variable Y as function of the variance of the covariate(s) weighted by the squared regression coefficient(s) [group-specific slopes]. Therefore, the model fulfills the requirement of accounting for heterogeneous between-group residual variances (see subsection 3.1.2).

Summary In this subsection we described the simple multi-group model as a first implementation of generalized analysis of covariance which is appropriate with respect to the implied variance structure. Nevertheless, for covariate-treatment regressions with interaction terms, the *ATE*–estimator cannot be expressed as a function of estimated model parameters. Therefore, the multi-group model proposed by Sörbom (1978) can only be applied to inferences about the average total effect if there is no interaction, and the extension for covariate-treatment regressions with interaction term suggested by Flory (2008) gives valid standard errors and test statistics only if the group size is considered as known number. If the group size is estimated from the data and the slope coefficients differ between groups, the uncertainty of estimating the mean of the treatment variable X as an estimate of the treatment probability $P(X = 1)$ has to be taken into account. In order to develop a generalized analysis of covariance usable for observational studies, the statistical model should not assume X as a fixed regressor.

In the next subsection we will introduce an elaborated multi-group model which includes the group size as an additional model parameter. Furthermore, we will discuss an augmentation approach for the multi-group model as suggested by Nagengast (2006) which can be applied to obtain more reliable standard errors for the *ATE*–estimator, provided that an additional assumption regarding the asymptotic variances of parameter estimates is fulfilled.

3.3.3.2 Elaborated Multi-Group Model (based on the `KNOWNCLASS`–Option)

The *elaborated multi-group model* is derived as an extension of the multi-group model presented in the previous subsection [see the path diagram in Figure 3.1]. Therefore, the presentation here is restricted to the different treatment of the group size.

Model Specification Based on the general framework for latent variable modeling (see B. O. Muthén, 2002), the group-specific regressions $E_{X=0}(Y|Z)$ and $E_{X=1}(Y|Z)$ are specified within the mixture modeling capabilities of `Mplus`. Mixture modeling allows the specification of a categorical variable C, which is technically a dichotomous latent variable, representing the observed group membership X (B. Muthén, 1998-2004). This extension gives us access to the treatment probability

$$P(X=0) = \frac{1}{\left(1+\exp[-E(C)]\right)}, \tag{3.56}$$

with $E(C)$ as the expectation of the categorical latent variable C. If we insert Equation (3.56) in Equation (3.52), we can express the unconditional expectation of the covariate $E(Z)$ as a function of model parameters because the model contains a parameter for the estimated $E(C)$:

$$E(Z) = E_{X=0}(Z) \cdot \frac{1}{\left(1+\exp[-E(C)]\right)} + E_{X=1}(Z) \cdot \left(1 - \frac{1}{\left(1+\exp[-E(C)]\right)}\right) = f(\boldsymbol{\theta}). \tag{3.57}$$

Listing 3.2 presents the syntax for the this model[25] as we will use it in the simulation study.

```
DATA:     FILE = example.dat;
VARIABLE: NAMES ARE Y X Z;
          USEVARIABLES ARE Z Y;
          CLASSES = c (2); KNOWNCLASS = c (X=0 X=1);
ANALYSIS: TYPE IS MIXTURE;
MODEL:    %OVERALL%
                   Y ON Z;
                   [c#1] (C);
          %c#1%
                   Y ON Z(beta01);        ! E(Y|Z,X=0)
                   [Y](beta00);           ! (Control)
                   [Z](EzX0);
                   Y; Z;
          %c#2%
                   Y ON Z(beta11);        ! E(Y|Z,X=1)
                   [Y](beta10);           ! (Treatment)
                   [Z](EzX1);
                   Y; Z;

MODEL TEST:  0 = (beta10 - beta00)
                 + beta11*( EzX0*(1/(1+exp(-C)))
                          + EzX1*(1-(1/(1+exp(-C)))) )
                 - beta01*( EzX0*(1/(1+exp(-C)))
                          + EzX1*(1-(1/(1+exp(-C)))) );
```

Listing 3.2: `Mplus`–syntax for the elaborated multi-group model

[25] According to the `Mplus`–syntax of listing 3.2, the following eleven terms are estimated as elements of θ in the specified order:

$$\theta = \left(\beta_{00} \quad E_{X=0}(Z) \quad \beta_{01} \quad Var(\varepsilon_{X=0}) \quad Var_{X=0}(Z) \quad \beta_{10} \quad E_{X=1}(Z) \quad \beta_{11} \quad Var(\varepsilon_{X=1}) \quad Var_{X=1}(Z) \quad E(C) \right)^T. \tag{3.58}$$

Standard Error for the Average Total Effect Unfortunately, the current version of `Mplus` (Version 5) does not compute the standard error of the nonlinear constraint as a "new parameter" by itself (due to restrictions within the mixture modeling framework under certain conditions). While lines 20-24 in Listing 3.2 implement the Wald–test for the test of the average total effect against zero, the computation of the standard error is currently not available.

As described in section 3.3.2, if a new parameter can be expressed as a smooth function of model parameters, $f(\boldsymbol{\theta})$, the δ-method can be applied for the computation of the standard error of the new parameter. Inserting Equation (3.57) into $ATE_{10} = (\beta_{10} - \beta_{00}) + (\beta_{11} - \beta_{01})E(Z)$ gives the nonlinear constraint as:

$$\begin{aligned} c_{ATE_{10}}(\boldsymbol{\theta}) = \quad & (\beta_{10} - \beta_{00}) \\ & + \beta_{11} \cdot \left[E_{X=0}(Z) \cdot \left(\tfrac{1}{1+\exp[-E(C)]} \right) + E_{X=1}(Z) \cdot \left(1 - \tfrac{1}{1+\exp[-E(C)]} \right) \right] \\ & - \beta_{01} \cdot \left[E_{X=0}(Z) \cdot \left(\tfrac{1}{1+\exp[-E(C)]} \right) + E_{X=1}(Z) \cdot \left(1 - \tfrac{1}{1+\exp[-E(C)]} \right) \right]. \end{aligned} \tag{3.59}$$

The square root of the asymptotic variance of the constraint of Equation (3.59) equals the standard error of the *ATE*–estimator. The formula for computing the asymptotic variance of $f(\hat{\boldsymbol{\theta}}) = c_{ATE_{10}}(\hat{\boldsymbol{\theta}})$ is given in Equation (3.48). To apply this method, we derived the partial derivatives of the constraint $c_{ATE_{10}}(\hat{\theta})$ with respect to the $K = 11$ elements of the parameter vector θ, i. e., $\hat{\mathbf{C}} = \dfrac{\partial c(\hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}^T}$. Based on the vector $\boldsymbol{\theta}$ of the $K = 11$ model parameters for the model with two groups and one covariate (see footnote 25), we obtained the following partial derivatives:

$$\hat{\boldsymbol{C}} = \begin{pmatrix} -1 \\ \beta_{11} \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) - \beta_{01} \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) \\ -E_{X=0}(Z) \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) - E_{X=1}(Z) \cdot \left(1 - \frac{1}{1+(\exp[-E(C)])} \right) \\ 0 \\ 0 \\ 1 \\ \beta_{11} \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) - \beta_{01} \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) \\ E_{X=0}(Z) \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) + E_{X=1}(Z) \cdot \left(1 - \frac{1}{1+(\exp[-E(C)])} \right) \\ 0 \\ 0 \\ \begin{pmatrix} \beta_{11} \cdot \left[E_{X=0}(Z) \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) + E_{X=1}(Z) \cdot \left(1 - \frac{1}{1+(\exp[-E(C)])} \right) \right] \\ -\beta_{01} \cdot \left[E_{X=0}(Z) \cdot \left(\frac{1}{1+(\exp[-E(C)])} \right) + E_{X=1}(Z) \cdot \left(1 - \frac{1}{1+(\exp[-E(C)])} \right) \right] \end{pmatrix} \end{pmatrix}. \tag{3.60}$$

The Wald–test requested at the end of the Mplus–syntax in Listing 3.2 is computed by the inverse of the estimated asymptotic covariance matrix of the constraint, pre- and postmultiplied by the value of the constraint evaluated at the estimated parameters [see Equation (3.43) or Bollen, 1989, p. 294].

3.3.3.3 Approximated Multi-Group Model

We shall now briefly describe the realization of the approach based on the augmented asymptotic variance-covariance matrix of parameter estimates developed by Nagengast (2006). The approach is based on the multi-group structural equation model as described in subsection 3.3.3.1 (see Figure 3.1 for the path diagram and Listing 3.1 for the syntax, of course without the faulty constraint in line 16-21).

Note that this method is also currently implemented in the software package *EffectLite* and is therefore described in detail in the manual (see Steyer & Partchev, 2008).[26] For convenience we will use the R package LACE (Steyer & Partchev, 2008) for the simulation study, which simplifies the implementation (see Listing 3.3).

```
# mplus.executable <- path to the mplus executable
#
# x  <- group membership
# y  <- outcome variable
# z  <- covariate


res <- lace(x,y,z,control.group="0",engine="mplus",
            program = mplus.executable)

# Results:
#
#   res$wald["Effect","Statistic"]
#   res$wald["Effect","p-value"]
```

Listing 3.3: R syntax for the approximated multi-group model based on LACE

Augmented Variance-Covariance-Matrix of Parameter Estimates The motivation for *augmenting the variance-covariance matrix of parameter estimates* comes from the observation that the relative group size (that is the mean of the treatment variable X) is not an estimated parameter in the multi-group structural equation model. Consequently, we arrive at no standard error and no asymptotic covariances of this estimated quantity with any other parameter of the model. As described in section 3.1.3, we assume X to be a stochastic regressor. Hence, the standard error of the average total effect should account for the uncertainty of the estimated treatment probability.

[26] In contrast to the elaborated multi-group model, generalized analysis of covariance based on the augmentation approach can be technically performed with all software packages for structural equation models with the multi-group option (e. g., LISREL, Jöreskog & Sörbom, 1988, or EQS, Bentler, 1995). The only requirement is that the estimated matrix of asymptotic variances and covariances is accessible, which is necessary for the augmentation approach.

The core idea of Nagengast (2006) on how to solve this problem is to compute the variance-covariance matrix of the assignment probabilities $acov(\mathbf{P})$ and insert it in the asymptotic variance-covariance matrix of parameter estimates used to apply the δ-method:

$$\mathbf{\Sigma}_{aug} = \begin{pmatrix} acov(\boldsymbol{\theta}) & \\ acov(\boldsymbol{\theta}, \mathbf{P}) & acov(\mathbf{P}) \end{pmatrix}. \tag{3.61}$$

The resulting matrix $\mathbf{\Sigma}_{aug}$ is labeled as *augmented variance-covariance matrix*, a $(K+J) \times (K+J)$ symmetric matrix. We can distinguish three parts:

For the first part $acov(\boldsymbol{\theta})$ of the matrix in Equation (3.61), the $K \times K$ matrix of asymptotic variances and covariance of the model parameters $\boldsymbol{\theta}$ estimated by minimizing F_{ML}, we can use the usual estimates [computed by Equation (3.47)].

For the second part $acov(\mathbf{P})$ in the diagonal of the matrix in Equation (3.61), the asymptotic variance-covariance matrix of the group sizes, Nagengast (2006) provided formulas to obtain a sample estimate. For a total of J groups, he created indicator variables for the group membership, $I_{X=j}$ (which sum up to one, i. e., $\sum_j \left[E(I_{X=j})\right] = 1$ with $j = 0, \ldots, J-1$, see footnote 21 for the definition of $I_{X=j}$). The expectation of these indicator variables are estimated by the relative frequencies as the sample estimates of the group sizes. Nagengast then focused the distribution of the matrix $\mathbf{I_X}$, which contains the J indicator variables $I_{X=j}$ and formalized the $J \times 1$ column vector of expectations and the $J \times J$ matrix of variances and covariances

$$E(\mathbf{I_X}) = \begin{pmatrix} P(X=1) \\ P(X=2) \\ \vdots \\ P(X=J) \end{pmatrix} \quad \text{and} \quad \Sigma(\mathbf{I_X}) = E\left(\mathbf{I_X}^2\right)\mathbf{1}^T - E\left(\mathbf{I_X}\right)E\left(\mathbf{I_X}\right)^T, \tag{3.62}$$

where $\mathbf{1}$ is a $J \times 1$ unit column vector. The vector $E(\mathbf{I_X})$ is equal to the vector of assignment probabilities, therefore the matrix $\mathbf{I_X}$ can be transformed to the variance-covariance matrix of group sizes:

$$acov(\mathbf{P}) = \frac{\Sigma(\mathbf{I_X})}{N-1}. \tag{3.63}$$

To obtain an unbiased estimator of the variance-covariance matrix $\Sigma(\mathbf{I_X})$, the expectations of the indicator variables $E(\mathbf{I_X})$ are replaced by the estimated sample means $\hat{\boldsymbol{\mu}}(\boldsymbol{I_X})$. Moreover, because of dichotomous indicator variables $\hat{\boldsymbol{\mu}}(\mathbf{I_X}) = \hat{\boldsymbol{\mu}}(\mathbf{I_X}^2)$, Nagengast finally gives the following formula to estimate $acov(\mathbf{P})$:

$$\widehat{acov}(\mathbf{P}) = \hat{\Sigma}(\mathbf{I_X}) = \left[\hat{\boldsymbol{\mu}}(\mathbf{I_X})\mathbf{1}^T - \hat{\boldsymbol{\mu}}(\mathbf{I_X})\hat{\boldsymbol{\mu}}(\mathbf{I_X})^T\right]\frac{N}{(N-1)^2}. \tag{3.64}$$

For the simple case of two group ($J = 2$) considered in this thesis, we can reduce the formula to

$$\widehat{acov}(\mathbf{P}) = \left[\hat{\mu}(X) - \hat{\mu}(X)^2\right] \frac{N}{(N-1)^2}. \tag{3.65}$$

Nevertheless, for the third part $acov(\boldsymbol{\theta})$ of the augmented variance-covariance matrix in Equation (3.61), Nagengast (2006) assumes that the group sizes are uncorrelated with the other model parameters, i. e., the following assumption is necessary:

$$acov(\boldsymbol{\theta}, \mathbf{P}) = \mathbf{0}. \tag{3.66}$$

As a result, we rely on this assumption about the asymptotic covariances between the group size $\mathbf{P}$ and the model parameters $\boldsymbol{\theta}$ for the application of the δ-method. If this assumption is valid, the augmented variance-covariance matrix of parameter estimates can be used to derive a standard error for the average total effect [see Equation (3.48)] and to conduct a Wald–test (see paragraph 3.3.2).

To compute the standard error of the *ATE*–estimator and to apply the Wald–test, the estimator of the average total effect is considered as a function of $\boldsymbol{\theta}$ and $P(X = 1)$ in *EffectLite* and `LACE`:

$$\begin{aligned} c_{ATE_{10}}\big(\boldsymbol{\theta}, P(X=1)\big) = \quad & \big(\beta_{10} - \beta_{00}\big) \\ & + \big(\beta_{11} - \beta_{01}\big) \cdot \big[E_{X=1}(Z) \cdot P(X=1) + E_{X=0}(Z) \cdot \big(1 - P(X=1)\big)\big]. \end{aligned} \tag{3.67}$$

With the help of the partial derivatives for $\mathbf{C}_{aug} = \dfrac{\partial c_{ATE_{10}}\big(\boldsymbol{\theta}, P(X=1)\big)}{\partial \boldsymbol{\theta}^T}$,[27] and under the assumption that $acov(\boldsymbol{\theta}, \mathbf{P}) = 0$, the estimate of the vector of partial derivatives $\widehat{\boldsymbol{C}}_{aug}$ and the estimated augmented variance-covariance matrix of parameter estimates $\widehat{\boldsymbol{\Sigma}}_{aug}$ are used to compute the standard error $S.E.\big(\widehat{ATE}_{10}\big) = \sqrt{\widehat{avar}\big[c_{ATE_{10}}(\widehat{\boldsymbol{\theta}}, \widehat{\mu}_{I_{X=1}})\big]}$ of the *ATE*–estimator based on the asymptotic variance:

$$\widehat{avar}\big[c_{ATE_{10}}(\widehat{\boldsymbol{\theta}}, \widehat{\mu}_{I_{X=1}})\big] = \hat{\mathbf{C}}_{aug}\, \Sigma_{aug}\, \hat{\mathbf{C}}_{aug}^T. \tag{3.69}$$

[27] See footnote 23 for the $K = 10$ elements of the vector $\boldsymbol{\theta}$.

$$\boldsymbol{C}_{aug} = \begin{pmatrix} -1 \\ \big(\beta_{11} - \beta_{01}\big) \cdot (1 - P(X=1)) \\ P(X=1) \cdot \big(E_{X=0}(Z) - E_{X=1}(Z)\big) - E_{X=0}(Z) \\ 0 \\ 0 \\ 1 \\ \big(\beta_{11} - \beta_{01}\big) \cdot P(X=1) \\ P(X=1) \cdot \big(E_{X=1}(Z) - E_{X=0}(Z)\big) + E_{X=0}(Z) \\ 0 \\ 0 \\ \big(E_{X=1}(Z) - E_{X=0}(Z)\big) \cdot \big(\beta_{11} - \beta_{01}\big) \end{pmatrix}. \tag{3.68}$$

The Wald–test printed by `LACE` (and *EffectLite*) is computed as:

$$\text{Wald} = \left(c_{ATE_{10}}(\hat{\boldsymbol{\theta}}, \hat{\mu}_{I_{X=1}})\right)^T \left(\widehat{avar}\left[c_{ATE_{10}}(\hat{\boldsymbol{\theta}}, \hat{\mu}_{I_{X=1}})\right]\right)^{-1} \left(c_{ATE_{10}}(\hat{\boldsymbol{\theta}}, \hat{\mu}_{I_{X=1}})\right). \tag{3.70}$$

Transforming the Variance-Covariance Matrix of Parameter Estimates Testing the average total effect based on the *augmentation* of the asymptotic variance-covariance matrix of parameter estimates rests on the validity of the underlying assumption in Equation (3.66). With the help of the elaborated multi-group model described in the previous subsection, it is possible to study the plausibility of the assumption underlying the augmentation approach. For this purpose, the estimated asymptotic covariances between the estimated expectation of the categorical latent variables $E(C)$ and the remaining model parameters in $\boldsymbol{\theta}$ can be transformed into asymptotic covariances between the group size and the model parameters in the constraint, i. e., into $acov(\boldsymbol{P}, \boldsymbol{\theta})$. This transformation is possible again with the help of the multivariate δ-method. Unlike the application of the δ-method presented so far for the computation of the standard error for the average total effect, a matrix is used as the *Jacobian*. For the simple model with one covariate and two groups, this Jacobian is an 11×12 matrix, composed of an 11×11 diagonal identity matrix with one additional column. This additional column contains the derivation of the group size $P(X = 1)$ with respect to $E(C)$, the expectation of the latent categorical variable representing the group membership. The transformed asymptotic covariances are then obtained by pre- and postmultiplication of the estimated variance-covariance matrix of parameter estimates computed by `Mplus` with this 11×12 matrix of second derivatives.

Summary Three multi-group models for an implementation of generalized analysis of covariance are described in this thesis. The approximated multi-group model presented in this subsection takes the stochasticity of the covariate(s) into account and is free from the assumption of homogenous between-group residual variances. Only the assumption of uncorrelated parameter estimates is necessary for the augmentation. With the help of the Monte Carlo simulation and based on the elaborated multi-group model we will study the empirical covariation of the parameter estimates as well as the distribution of the estimated asymptotic covariances.

3.3.4 Specification of Single Group Structural Equation Models

In addition to the three multi-group models for an implementation of generalized analysis of covariance described so far, two single group approaches are presented in this subsection.

3.3.4.1 Simple Single Group Model (with Interaction)

Model Specification The implementation of generalized analysis of covariance as a *simple single group model* is for observed covariates comparable to the moderated regression model described in subsection 2.2.3. The main distinction is the different estimation method of the covariate-treatment regression as structural equation model and the different test statistic used to test hypotheses about the average total effect. A path diagram for this structural equation model as discussed by Flory (2008) is given in Figure 3.2. The intercept γ_{00} is not included in the path diagram and the black dot represents the interaction between X and Z (i. e., the product term $X \cdot Z$ defined as a separate predictor variable) with the associated regression coefficient γ_{11}.

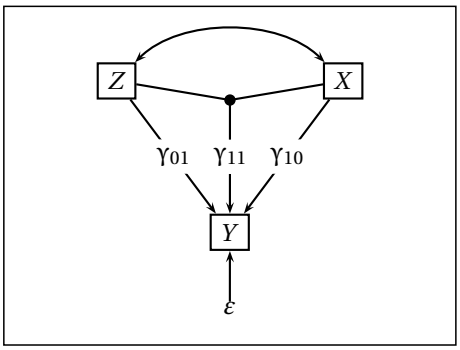


Figure 3.2: Path diagram of the simple single group model (with interaction)

Based on the corresponding `Mplus`–syntax (presented in Listing 3.4), the estimation of the model parameters is conducted with the help of full-information maximum likelihood. The *nonlinear constraint* for the average total effect, constructed by taking the expectation of the effect function $g_1(Z)$ [or similarly, by formulating the difference between the adjusted means as a function of model parameters], is equivalent to the formula presented for the moderated regression [see Equation (2.4) on page 27]. The corresponding constraint of the estimated model parameters is included in the syntax in lines 17 and 18, with the names `gamma10` for $\hat{\gamma}_{10}$, `gamma11` for $\hat{\gamma}_{11}$, and `Ez` for $\hat{\mu}_Z$. Note that this estimator itself is identical to the well-known *average distance* [see Equation (3.20) on page 64]. Nevertheless, the Wald–test requested with the command in line 20 of Listing 3.4 incorporates the stochastic nature of X and Z. As a result, `Mplus` computes the standard error for the *ATE*–estimator, which is internally computed by the δ-method as described in section 3.3.2, and prints a p-value for the hypothesis $H_0 : ATE_{10} = 0$.

For convenience, only the simple regression of one (observed) outcome variable Y on one (observed) covariate Z is described here. For real data applications the model can be further generalized to multivariate covariates $Z = (Z_1, ..., Z_K)$ or to latent covariates by adding an appropriate measurement model (see, Flory, 2008, for a formulation of this simple single group model with latent covariates). Note that with

```
DATA:     FILE = example.dat;
VARIABLE: NAMES ARE Y, X, Z;
          USEVARIABLES ARE Y, X, Z, ZX;

DEFINE:   ZX = Z*X;             ! Define the Predictor for the
                                ! Covariate-Treatment Interaction

ANALYSIS: TYPE = MEANSTRUCTURE;

MODEL:    Y ON Z(gamma01),      ! E(Y|X,Z)
             X(gamma10),
            ZX(gamma11);
           [Y](gamma00);
           [Z](Ez);             ! E(Z)
          [X ZX];

MODEL CONSTRAINT: new(ATE10);
                  ATE10 = gamma10 + gamma11 * Ez;

MODEL TEST:       0 = ATE10;
```

Listing 3.4: `Mplus`–syntax for the simple single group model (with interaction)

an interaction between a latent continuous covariate and an observed variable, the simple single group model is estimated by a different estimator than the maximum likelihood estimator described above (see B. O. Muthén & Asparouhov, 2003, and see also the elaborated single group model described in the next subsection). If multiple covariates are used, the model is probably more complicated to specify because all interactions have to be included. For multiple group comparisons ($J > 2$), the model can be extended by introducing (further) indicator variables indicating group membership.

Implied Variance Structure In contrast to the multi-group models discussed in the previous subsection, the simple single group model with interaction specifies one single residual term ε for the covariate-treatment regression (see also Figure 3.2). As we have derived in section 3.1.2 from the theory of stochastic causality, we expect *heterogeneity* of residual variances between treatment groups, that is with respect to the implied variance structure we require that the model accounts for $Var_{X=1}(\varepsilon) \neq Var_{X=0}(\varepsilon)$. For an illustration of this *misspecification* of the simple single group model, it is revealing to consider the implied variance of the outcome variable Y in the treatment condition $X = 1$,

$$\begin{aligned} Var_{X=1}(Y) &= Var\left(\gamma_{00} + \gamma_{01}Z + \gamma_{10}X + \gamma_{11}ZX + \varepsilon\right) \\ &= Var\left((\gamma_{00} + \gamma_{10}) + (\gamma_{01} + \gamma_{11})Z + \varepsilon\right) \\ &= (\gamma_{01} + \gamma_{11})^2 Var(Z) + Var(\varepsilon), \end{aligned} \tag{3.71}$$

and in the control condition $X = 0$,

$$\begin{aligned} Var_{X=0}(Y) &= Var\left(\gamma_{00} + \gamma_{01}Z + \gamma_{10}X + \gamma_{11}ZX + \varepsilon\right) \\ &= Var\left(\gamma_{00} + \gamma_{01}Z + \varepsilon\right) \\ &= \gamma_{01}^2 Var(Z) + Var(\varepsilon). \end{aligned} \tag{3.72}$$

For the case of no covariate-treatment interaction ($\gamma_{11} = 0$) the implied variance is equal between treatment groups, i. e., $Var_{X=j}(Y) = \gamma_{01}^2 Var(Z) + Var(\varepsilon)$ for $j = 1$ and $j = 0$. Nevertheless, we did not assume an interaction for deriving the implications of heterogeneity of residual variances in section 3.1.2. The heterogeneity of residual variances is already a consequence of individual and heterogeneous treatment effects. Moreover, the comparison of Equation (3.71) and Equation (3.72) reveals that different variances for the outcome variable Y conditional on X are implied only as a function of the estimated regression coefficient for the interaction term (γ_{11}).

For structural equation models, the parameters are estimated by simultaneously minimizing the difference between the observed variance-covariance matrix $\mathbf{S}_g$ and the implied variance-covariance matrix $\Sigma_g(\boldsymbol{\theta})$, as well as the squared difference between the implied mean vector $\overline{\boldsymbol{\nu}}_g$ and the observed mean structure $\boldsymbol{\mu}_g$ [see, e. g., Equation (3.39)]. In contrast to the general linear model, where heteroscedasticity is known to have an effect on the estimated standard error, the *specification error* of structural equation models might also bias the estimated parameters. Therefore, it is reasonable to expect consequences of this misspecification in the implied variance structure on parameter estimates of the regression coefficients obtained from the simple single group model (see subsection 3.3.1).

Summary In this section we have described the simple single group model. For this model, the stochasticity of X is not relevant because the model is not estimated conditional on the group membership, and instead the estimate of the unconditional expectation of the covariate $E(Z)$ is included in the nonlinear constraint. As we have described, the implied variance structure of this model is not valid with respect to the heterogenous between-group residual variances implied by the theory of stochastic causality. Therefore, a similar issue as discussed for the general linear model with respect to heterogeneity of residual variances applies. Nevertheless, the simple single group model, suggested by Flory (2008), represents an improvement compared to the approaches based on the general linear hypothesis. This suggestion was corroborated by the derivation that for valid unconditional inference about the average total effect the joint distribution of Y, X and Z must be taken into account. In the next subsection we will present an extension of the single group model that in addition incorporates heterogenous residual variances.

3.3.4.2 Elaborated Single Group Model (with Random Slope)

Over the last years, new methods for estimating nonlinear structural equation models with interaction terms have been developed (see, e. g., Schumacker & Marcoulides, 1998; Schumacker, 2002; Hancock & Mueller, 2006; S.-Y. Lee, 2007). A highly influential approach was suggested by Kenny and Judd (1984), and extended and modified by Jöreskog and Yang (1996). To give an extensive review of the literature is beyond the scope of this section. For a review of the performance of the different methods see, for example, Marsh, Wen, and Hau (2004).

Recently, based on the work of Klein and Moosbrugger (2000), a very promising estimation method (*latent moderated structural equations*, `LMS`–estimator) has become widely available through the implementation in `Mplus` (see also Klein & Muthén, in press). As B. O. Muthén and Asparouhov (2003) describe, interactions with latent continuous variables can be rewritten as two equations involving a *random slope* variable. Treating X as a continuous observed variable, we combine this approach with the suggestion given by B. O. Muthén and Asparouhov (2002) for the modeling of heteroscedastic residual variances. This leads to an implementation of generalized analysis of covariance as a single group structural equation model which can be estimated with the `LMS`–approach under the assumption of normality of the (latent) predictor variables and the residual terms (see, e. g., Moosbrugger et al., in press). This approach is similar to the random slopes multi-level structural equation models for heteroskedasticity of known form (see, e. g., Rabe-Hesketh, Skrondal, & Pickles, 2004). We will therefore call the resulting model *elaborated single group model with random slope*.

Model Specification To describe the model specification of generalized analysis of covariance with interaction terms as elaborated single group model, we start again with the classical ANCOVA model:

$$\begin{aligned} E(Y|X,Z) &= (\gamma_{00} + \gamma_{01} \cdot Z) + \gamma_{RS} \cdot X \\ \varepsilon &\equiv Y - E(Y|X,Z). \end{aligned} \tag{3.73}$$

Unlike the simple single group model, we do not simply introduce an additional predictor $X \cdot Z$ for the covariate-treatment interaction into the regression. Rather we assume the regression coefficient γ_{RS} for the linear regression Y on X to be a *random slope*.[28] This random slope itself is regressed on the covariate(s):

$$\begin{aligned} E(\gamma_{RS}|Z) &= \gamma_{10} + \gamma_{11} \cdot Z \\ \varepsilon_{\gamma_{RS}|Z} &\equiv \gamma_{RS} - E(\gamma_{RS}|Z), \end{aligned} \tag{3.74}$$

[28] Although Flory (2008) called his single group model *random slope*, this term is used within this thesis for the model where the slope of the regression Y on Z is regressed itself on X, with an additional residuum not restricted to zero variance.

with the assumption that $\varepsilon_{\gamma_{RS}|Z}$ and ε are uncorrelated (B. O. Muthén & Asparouhov, 2002).

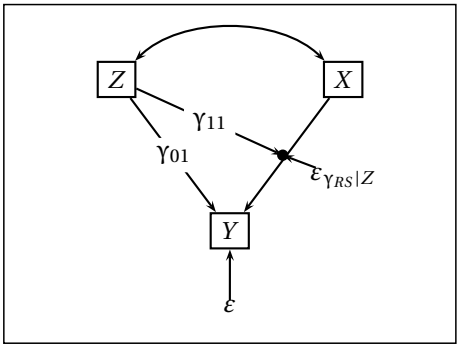


Figure 3.3: Path diagram of the elaborated single group model

The resulting path diagram is shown in Figure 3.3. Again, as for the simple single group model with interaction, a special notation is necessary to represent the interaction between X and Z. Apart from the path diagram given in Figure 3.2, the dot in Figure 3.3 represents the random slope γ_{RS} that is regressed on Z [see the corresponding regression coefficient γ_{11} and the residual term $\varepsilon_{\gamma_{RS}|Z}$, which correspond to the residual in Equation (3.74)].

For technical reasons, the corresponding `Mplus`–syntax presented in Listing 3.5 is slightly more complicated because a dummy variable approach is necessary for using the random slope approach with observed covariates. If the approach is applied with latent covariates, i. e., if an appropriate measurement model is specified for covariates measured with measurement error, the dummy variable is not necessary as it will be replaced by the latent covariate. As indicated by Steyer et al. (in press), a linear measurement model will not change the general modeling strategy for generalized analysis of covariance.

Putting Equation (3.73) and Equation (3.74) together leads to the same parameterization as the *simple single group model* for the regression coefficients, and for the implied mean structure:

$$\begin{aligned} E(Y|X,Z) &= g_0(Z) + g_1(Z)\cdot X \\ &= \left(\gamma_{00} + \gamma_{01}Z\right) + \gamma_{RS}\cdot X \\ &= \left(\gamma_{00} + \gamma_{01}Z\right) + \left(\gamma_{10} + \gamma_{11}Z + \varepsilon_{\gamma_{RS}|Z}\right)\cdot X. \end{aligned} \tag{3.75}$$

Furthermore, because the expectation of the random slope residual $E(\varepsilon_{\gamma_{RS}|Z})$ is zero by definition, the constraint for the average total effect remains unchanged:

$$\begin{aligned} ATE_{10} &= E\left(g_1(Z)\right) \\ &= E\left(\gamma_{10} + \gamma_{11}Z + \varepsilon_{\gamma_{RS}|Z}\right) \\ &= \gamma_{10} + \gamma_{11}\cdot E(Z). \end{aligned} \tag{3.76}$$

```
DATA:     FILE = example.dat;                                          1
VARIABLE: NAMES ARE Y, X, Z;                                           2
ANALYSIS: TYPE = RANDOM;                                               3
          ALGORITHM = INTEGRATION;                                     4
                                                                       5
MODEL:    dummy_z BY z@1;              ! Define dummy covariate as latent  6
          [z@0];                       ! (necessary to apply the random    7
          z@0;                         !  slope approach for manifest      8
                                       !  variables, i. e., dummy_z := z)  9
          x WITH dummy_z;                                              10
          slope | y ON x;              ! E(Y|X,Z)                      11
          y ON dummy_z;                                                12
                                                                       13
          slope ON dummy_z (gamma11); ! E(gamma*|Z)                    14
                                                                       15
          slope(epsilon);              ! Var(epsilon)                  16
          [dummy_z](Ez);               ! E(Z)                          17
          [slope](gamma10);            !                               18
                                                                       19
MODEL CONSTRAINT: new(ATE10);                                          20
                  ATE10 = gamma10 + gamma11 * Ez;                      21
                                                                       22
MODEL TEST:       0 = ATE10;                                           23
```

Listing 3.5: `Mplus`–syntax for the elaborated single group model

All parameters necessary to compute the nonlinear constraint are part of the estimated parameters of the `Mplus`–model ($\widehat{\boldsymbol{\theta}}$). We can make use of the built-in capability of `Mplus` to formulate the estimator of the average total effect as a "new parameter" [see that lines 20-21 of Listing 3.5 match the constraint given in Equation (3.76)]. Finally, the Wald–test for the hypothesis $ATE_{10} = 0$ is requested by the command in line 23 of the listing.

Implied Variance Structure The main difference between the simple and the elaborated single group model becomes clear when taking a closer look at the implied variance of the elaborated single group model for the outcome variable in the treatment condition $X = 1$ [derived from Equation (3.75)], i. e.,

$$\begin{aligned} Var_{X=1}(Y) &= Var\left[\left(\gamma_{00}+\gamma_{01}Z\right)+\left(\gamma_{10}+\gamma_{11}Z+\varepsilon_{\gamma_{RS}|Z}\right)\cdot X+\varepsilon\right] \\ &= Var\left[\gamma_{00}+\gamma_{01}Z+\gamma_{10}+\gamma_{11}Z+\varepsilon_{\gamma_{RS}|Z}+\varepsilon\right] \\ &= Var\left[(\gamma_{00}+\gamma_{10})+(\gamma_{01}+\gamma_{11})Z+\varepsilon_{\gamma_{RS}|Z}+\varepsilon\right] \\ &= (\gamma_{01}+\gamma_{11})^2\,Var(Z)+Var(\varepsilon)+Var(\varepsilon_{\gamma_{RS}|Z}), \end{aligned} \tag{3.77}$$

and the implied variance for the outcome variable in the control condition $X = 0$, i. e.,

$$\begin{aligned} Var_{X=0}(Y) &= Var\left[\left(\gamma_{00}+\gamma_{01}Z\right)+\left(\gamma_{10}+\gamma_{11}Z+\varepsilon_{\gamma_{RS}|Z}\right)\cdot X+\varepsilon\right] \\ &= Var\left(\gamma_{00}+\gamma_{01}Z+\varepsilon\right) \\ &= \gamma_{01}^2\,Var(Z)+Var(\varepsilon). \end{aligned} \tag{3.78}$$

To derive Equation (3.77) we make use of the property that $Cov(\varepsilon_{\gamma_{RS}|Z}, \varepsilon) = 0$. If we now assume that there is no interaction ($\gamma_{11} = 0$), we see that the elaborated single group model implies a variance structure with different variances for the two treatment groups. The difference equals the variance of the random slope residual [i. e., $Var(\varepsilon_{\gamma_{RS}|Z})$]. Under the more general condition with interaction ($\gamma_{11} \neq 0$), the implied variance structure differs in two parts between treatment and control groups: The variance due to the covariate-treatment interaction, that is $(\gamma_{01} + \gamma_{11})^2 Var(Z)$ versus $\gamma_{01}^2 Var(Z)$, and the variance of the random slope, i. e., the difference between-group residual variances which is not explained by the covariate, $Var(\varepsilon_{\gamma_{RS}|Z})$.

Summary To summarize the presentation of the elaborated single group model, we will verify that the model fulfills the three implications of the theory of stochastic causality we formulated for generalized analysis of covariance model: Firstly, the interaction between the covariate and the treatment variable (see section 3.1.1) is included by regressing the (random) slope of the regression of Y on X, i. e., γ_{RS}, on the covariate Z. Secondly, the stochasticity of the covariate (see section 3.2.5) is appropriately considered by implementing the model within the framework of structural equation modeling, as we obtain the parameters from maximum likelihood estimation (derived under the assumption of normally distributed residuals and predictors). The stochasticity of the group size is not relevant for the single group model, because for this implementation of generalized analysis of covariance we estimate the unconditional expectation of the covariate $E(Z)$ as a model parameter directly. Therefore, we "avoid" the difficulties introduced by weighting the conditional (i. e., group-specific) expectations of Z given $X = j$. Thirdly, the model implied variance structure of the elaborated single group model is correctly specified with respect to the heterogeneity of between-group residual variances (see section 3.1.2), as the regression slope γ_{RS} is assumed to be random (and itself regressed on the covariate). The additional residual term $\varepsilon_{\gamma_{RS}|Z}$ [i. e., $Var(\varepsilon_{\gamma_{RS}|Z})$ in Equation (3.77)] accounts for heterogeneity and makes the implied variance structure similar to the structure derived from theory [see Equation (3.4)].

The elaborated single group model can easily be used to consider measurement error of covariates and outcomes by the specification of a measurement model. Finally, the model can be generalized to the comparison of more than two treatment groups by adding additional indicator variables $I_{X=j}$ for each extra treatment group. For real data problems, this might again be cumbersome because for each group a random slope must be specified. As a consequence, each (necessary) covariate-treatment interaction must be included in the regression model.

3.4 Summary and Research Questions

The last part of this chapter presents a summary of research questions to be studied with the Monte Carlo simulation presented in chapter 4. Based on the theory of stochastic causality, we discussed implications

for the statistical model and found out that a test of the average total effect for quasi-experimental designs as the general linear hypothesis is not valid if interactions are present. We derived this conclusion as a consequence of the assumption of fixed regressors which lead to an underestimation of the variance of the *ATE*–estimator. Moreover, two misspecified structural equation models for generalized analysis of covariance were presented. Furthermore, we developed two structural equation models for the implementation of a generalized analysis of covariance, which are expected to give valid point estimates, standard errors, and test statistics for hypothesis about the average total effect. Moreover, we described an approximation, which is most general in the sense that it is applicable with a plurality of software packages for structural equation modeling. Hence, the remaining research questions can be subsumed into three categories: 1) An analysis of the robustness of the misspecified structural equation models; 2) A comparison of the performance of theoretical suitable methods for testing the average total effect resulting in recommendations for real data analysis, as well as 3) The derivation of further insights into the appropriateness of the necessary assumption for the augmentation approach of Nagengast (2006) and Steyer and Partchev (2008).

Robustness studies have a long tradition in statistical science (see, for example, Eye & Schuster, 1998, ch. 11 for *robustness of regression*). Ito (1980, p. 199) points out that a *"desirable characteristic of a test is that while it is powerful, i. e., sensitive to changes in the specified factors under test, it is robust, i. e., insensitive to changes in extraneous factors not under test. Specifically, a test is called robust when its significance level (Type-I error probability) and power (one minus Type-II error probability) are insensitive to departures from the assumptions on which it is derived."* Nevertheless, robustness cannot be studied globally because as noted already by Scheffe (1959), a study of the robustness properties of a test cannot be exhaustive, especially because the underlying assumptions can be violated in many more ways than they can be satisfied. This is important to keep in mind because beyond the three implications for the statistical model considered in this thesis, for example, the functional form assumption of the covariate-treatment regression is of major importance (see subsection 5.4.1).

Essentially, we will present results from a Monte Carlo simulation for which data were generated in line with a linear parameterized covariate-treatment regression (see section 4.2.2 for details). Therefore, all outcome regressions will be precisely correctly specified with respect to the adjusted means. It is expected that the ordinary least-squares estimates based on the observed values of the group size or based on the mean of the covariate should give unbiased estimates for the average total effect. Biases due to specification errors are not expected for ordinary least-squares regressions. For structural equation models, unbiased estimates are not expected if the implied variance structure is misspecified. Therefore, although the main outcome of the simulation study is a comparison of the performance of the different test statistics and standard errors, we will examine the absolute bias of the *ATE*–estimator as well.

Beyond these simple comparisons, the simulation study shall be conducted to illustrate and to study the two newly introduced structural equation models (the elaborated single group model and the elaborated multi-group model) in more detail. Before an elaborated (technical) description of the Monte Carlo simulation is given in chapter 4, we will present seven specific research questions in the next subsections.

3.4.1 Ignoring the Stochasticity of Z

To demonstrate the consequences of the analytic discussion of the unconditional variance of the average total effect estimator obtained from (conditional) ordinary least-squares regression (presented in section 3.2.5), the hypothesis of no average total effect will be tested for all generated datasets in the Monte Carlo simulation with the general linear hypothesis and with the mean-centering approach. Due to the violated requirements of the general linear model [see Equation (3.32) on page 69], underestimated standard errors and subsequently heavily inflated type-I-error rates are expected for conditions with non-parallel regression slopes. Mean-centering should provide exactly the same results as the general linear hypothesis if the estimated mean of the covariate is used.

For the sake of completeness, we will also analyze the data assuming that the expectation of the covariate is known and accessible to specify the linear hypothesis (or to transform the covariate in order to compute mean-centered covariates). It is expected that the ordinary least-squares variances of *ATE*–estimator will be unbiased when the true population value of the covariates mean is incorporated.

3.4.2 Robustness to Heterogeneity of Residual Variance

The robustness of the adjustment procedure with respect to heterogeneity of between-group residual variances is of interest for two different reasons: For approaches based on the general linear model due to heteroskedasticity, and for approaches based on structural equation modeling due to consequences of a possible misspecification of the implied variance structure.

For test statistics of the average total effect based on the general linear model (regardless of heteroskedasticity), we do not expect acceptable performance if interactions between baseline covariates and the treatment variable are present (if the sample estimates of the covariate's mean are used in the linear hypothesis or for the mean-centering approach). Nevertheless, an illustration of the two different sources for standard error biases, either due to the stochasticity of regressors, or due to heteroskedasticity for unequal group sizes, might lead to general statements about the robustness of the general linear model for testing average total effect for data obtained from quasi-experimental designs.

Furthermore, in order to study the corrected standard errors for regression estimates (Schafer & Kang, 2008) and the standard errors obtained by predictive simulations (Gelman & Hill, 2007), the performance of these approaches with respect to their robustness against heterogeneity of residual variances is of interest.

We will focus on performance of *heteroskedasticity consistent estimators* of the variance-covariance matrix of parameter estimates and compare the robust standard errors with the corrected standard errors obtained for the regression estimates and the standard errors estimated by predictive simulations. We will also analyze the small sample performance of the two selected robust estimators `HC3` (J. G. MacKinnon & White, 1985) and `HC4` (Cribari-Neto, 2004) in the Monte Carlo simulation. Therefore, the amount of between-group residual variance heterogeneity is one manipulated factor of the simulation design, as is the sample size (see section 4.3 for details).

The simple single group model (see section 3.3.4.1) was shown to be misspecified with respect to the implied variance structure, if we allow for heterogeneity of between-group gresidual variance as predicted by the theory of stochastic causality. Therefore, another goal of the Monte Carlo simulation is to study the robustness of the simple single group approach against this theoretical reasonable violation of the homogeneity of residual variances. This will lead to a comparison of the elaborated single group model with the simple single group model because we have developed this more elaborated single group alternative model exactly to fix the misspecified implied variance. In paragraph 3.2.2.1 (see page 56) we figured out that the robustness of methods based on the general linear model with respect to the assumption of homoskedasticity is limited for unequal group sizes. Therefore, the performance of the simple single group model under conditions where the treatment probability is different from $P(X = 1) = 0.5$ is of particular interest.

3.4.3 Accuracy of the Estimated Asymptotic Variance-Covariance Matrices

The accuracy of the estimated asymptotic variances and covariances of parameter estimates became prominent because the standard errors for the *ATE*–estimator for the proposed structural equation models are computed using the δ-method. As noted by D. P. MacKinnon (2008), *"standard errors based on the multivariate delta method are generally based on large sample theory and are best checked in a statistical simulation."* For a subset of the conditions investigated in our simulation study, Nagengast (2006) reported unbiased standard errors of the average total effect estimator based on the augmentation approach. Hence, we aim to replicate the findings of Nagengast (2006) with respect to the accuracy of the asymptotic variance-covariance matrices for the two elaborated structural equation models developed in this thesis. Furthermore, we shall attempt to generalize the finding that the standard error for the *ATE*–estimator based on different implementations of generalized analysis of covariance can be constructed with the help of the δ-method to conditions with heterogeneous between-group residual variances. After all, the performance of the standard error of the average total effect estimator based on the structural equation models has not yet been studied empirically in a simulation study when the true effect is different from zero.

The elaborated multi-group model is particularly of interest with respect to this research question because the elaborated multi-group model is estimated within the framework of mixture modeling in `Mplus`. Similarly, the accuracy of the asymptotic variance-covariance matrix of the elaborated single group model is considered because here the `LMS` maximum likelihood estimation is performed by numerical integration (see subsection 3.3.3.2). For both rather new estimation procedures, the empirical inspection of the variance-covariance matrix of parameter estimates is still pending for generalized analysis of covariance models.

3.4.4 Ignoring the Stochasticity of X

We pointed out that according to the premises of the δ-method, we expect the multi-group model to fail for stochastic regressors. An empirical verification of the prediction that the simple multi-group model does not give appropriate standard errors for the *ATE*–estimator is of interest as we cannot provide references beside Nagengast (2006), Steyer and Partchev (2008), and Flory (2008). To make matters worse, Flory (2008) and Nagengast (2006) reported contradictory results. Therefore, we (once again) analyze the robustness of the simple multi-group approach against the assumption of known group sizes in the Monte Carlo simulation conducted for this theses. To force clear results, we generate data with a wide range of parameters, even under extreme conditions. Analyzing the same data in addition with the two alternative multi-group models (i. e., the approximated multi-group model and the elaborated multi-group model) will give further insights, as described in the additional research question in the next subsection.

3.4.5 Assumption of Uncorrelated Parameter Estimates

The validity of conclusions derived from the approximated multi-group model is subject to the assumption that some particular parameter estimates are uncorrelated (see paragraph 3.3.3.3 on page 89). Beyond the overall performance of the approximated multi-group model (with respect to small sample sizes, convergence rate and statistical power) we will investigate the appropriateness of the assumption presented in Equation (3.66). Since there is no analytical proof known to us, the plausibility of this assumption is examined empirically with generated datasets.

Of course, the strength of results about asymptotic covariances obtained by a simulation study should be taken as preliminary evidence only because they might depend on the selected model for the data generation as well as on the particular values of the parameters involved in the data generation process. Nevertheless, if we can find substantial covariances between the group size estimate and other model parameters, or systematic differences in terms of biased standard errors between the elaborated multi-group model and the approximated multi-group model, this would at least uncover potentially unsolved problems.

3.4.6 Regression Estimate and Predictive Simulation

The standard errors for the regression estimate approach recently suggested by Schafer and Kang (2008) are described as robust to a misspecified implied variance structure (see paragraph 3.2.2.1, on page 58). Within the Monte Carlo simulation we will study the self-evident question of how these standard errors perform for stochastic covariates. In other words, we will try to confirm their appropriateness for the test of the hypothesis of no average total effect for observational data when covariate-treatment interactions are present, and when the values of the covariates are observed rather than fixed by design.

With this question in mind, we will also apply the predictive simulation approach (see paragraph 3.2.3, on page 61) in order to find out whether or not this strategy can be used to overcome the restrictions of the ordinary least-squares regressions with respect to the underestimated variance of the average total effect estimator due to the nonlinearirty of the hypothesis.

3.4.7 Sample Size Requirements and Model Comparison

Finally, the central aim of a second part of the simulation study is a comparison of the two apparently correctly specified structural equation modeling approaches (the elaborated single group and the elaborated multi-group model) with the approximated multi-group model, which is already implemented in *EffectLite* (Steyer & Partchev, 2008). Here the most important question will be: How do the described models behave under conditions with small sample sizes?

To recommend any of the studied approaches for testing average total effects based on an outcome regression, the type-I-error rates are not sufficient. In addition to the (relative) bias of the standard error for generated data under a variety of conditions, we will also compare the statistical power to discover average total effects of the considered implementations of generalized analysis of covariance.

Chapter 4
Simulation Study

4.1 Introduction

In this chapter a Monte Carlo simulation with two parts is presented. Part I deals with the performance of the developed structural equation models compared to the methods based on ordinary least-squares estimated covariate-treatment regressions for data generated with no average total effect (see the research questions in subsections 3.4.1, 3.4.2, 3.4.3, 3.4.4 and 3.4.6). Part II of the simulation study is designed to analyze the statistical power and sample size requirements for the remaining subset of models that provided unbiased estimators of the average total effect, reasonable standard errors, and therefore correct type-I-error rates for the hypothesis $ATE = 0$ (see the research question in subsection 3.4.7).

This chapter will be structured as follows: The next section describes the data generation in detail. This is followed by a brief description of the simulation design and an introduction of the dependent measures. Thereafter, the results are presented in four sections. The first section demonstrates how the variance of the *ATE*–estimator obtained from applications of the general linear model is underestimated. The second result section discusses the performance of the structural equation models with nonlinear constraints under homogeneity of between-group residual variance. The following third result section presents the findings for the nonlinear constraint implemented in structural equation models for conditions with heterogeneity of between-group residual variance are presented. These three sections include the results from only part I of the simulation study. In the fourth result section, we focus on the model comparison based on part II of the simulation study. Each of the four result sections concludes with a summary of the most important findings. Finally, chapter 5 summarizes the results with respect to the seven research questions and presents general conclusions and recommendations for implementing a generalized analysis of covariances with interaction terms and stochastic regressors.

4.2 Data Generation

Datasets for the Monte Carlo study were generated in accordance with the theory of stochastic causality in order to study how different implementations of generalized analysis of covariance for quasi-experimental designs perform when the covariate Z and the treatment variable X are stochastic regressors.

Overview Datasets for the simulation study were generated in the following four steps: In the first step, a single univariate covariate Z was created according to the different sample sizes N, as summarized in

Table 4.1. A standardized *normally distributed covariate* Z^* with an expectation of zero and a variance equal to one was generated for each replication of the data generation (see Listing 4.1 for details), i. e., $Z^* \sim Norm(0,1)$. The simulated N subjects are assumed to be a random sample from an infinite universe (see Schochet, 2009, as well as section 1.1.9). In the second step, the values of the true outcome variable τ_0 were generated according to the parameterization of the intercept function, and the values of the individual total effect variable δ_{10} were generated in line with the parameterization of the effect function. Besides different regression coefficients for the outcome model, different true residual variances were incorporated to vary the amount of between-group residual variance heterogeneity. In the third step, allocation to the treatment conditions $X = 0$ or $X = 1$ was simulated based on individual treatment probabilities computed from an assignment model. Theses probabilities were obtained as a function of the values of the standardized covariate Z^* and the additional true parameters of the assignment model. Finally, in the fourth step, the appropriate value of the outcome variable Y was assigned for each simulated subject. According to the treatment variable X, either τ_0 or $\tau_0 + \delta_{10}$ was used for the computation of the outcome variable.

Table 4.1: Sample sizes used for data generation in simulations I and II

Simulation Study	Sample Sizes N
I	100, 250, 400, 1000
II	20, 30, 50, 75, 100, 150, 200, 250, 500, 1000

4.2.1 Assignment Model

The treatment assignment was randomized conditional on the covariate Z based on a model for (conditional) *treatment probabilities*, computed as logistic transformation of the standardized covariate Z^*. This transformation was parameterized with the two parameters α_0 and α_1 for the simple case of two groups and one covariate as considered in the simulation studies I and II. In order to hold the α_i parameter constant for the generation of different expectations of the observed covariate $E(Z)$ the covariate was used as standardized Z^*, i. e.,

$$P(X = 0|Z^*) = \frac{1}{1 + \exp\left(\alpha_0 + \alpha_1 \cdot Z^*\right)}. \tag{4.1}$$

The model-implied treatment probability $P(X = 1|Z^* = z) = 1 - P(X = 0|Z^* = z)$ was compared to draws from a uniformly distributed random variable for each simulated unit (see line 9 in Listing 4.1). This data generation procedure ensured that the treatment assignment was randomized given the value of Z^* (i. e., the treatment assignment was generated to be strongly ignorable given Z^*). According to this procedure the *correlation* between the treatment variable X and the covariate Z^* (as well as the correlation between X and the transformed covariate Z) depends on the α_1 parameter, for a given value of the α_0 parameter [see Equation (4.1)]. We describe this correlation by the index of determination $R^2_{Y|Z}$ for dichotomous variables

(Nagelkerke, 1991). Furthermore, the *group size* $P(X = 1)$ depends on the parameters α_0 and α_1 used for generating the true treatment probabilities. Table 4.2 presents the selected values of α_0 and α_1, the resulting correlations between X and Z, and the resulting group sizes $P(X = 1)$.[1]

Table 4.2: Data generation (assignment model) used in simulations I and II

Nagelkercke's Coefficient of Determination $R^2_{X\|Z}$	Group Size $P(X=1)$	Logistic α_0	Logistic α_1	Pearson Correlation $Cor(X,Z)$
0.75	0.2	-4.3	-4.9	-0.65
0.75	0.5	0	-4.2	-0.73
0.75	0.8	4.3	-4.9	-0.65
0.5	0.2	-2.4	-2.3	-0.55
0.5	0.5	0	-2	-0.6
0.5	0.8	2.4	-2.3	-0.55
0.25	0.2	-1.8	-1.2	-0.39
0.25	0.5	0	-1.1	-0.44
0.25	0.8	1.8	-1.2	-0.39
0.1	0.2	-1.4	-0.7	-0.27
0.1	0.5	0	-0.6	-0.28
0.1	0.8	1.4	-0.7	-0.27

Note: Coefficient of determination is Nagelkercke $R^2_{X|Z}$. Correlation between X and Z is given as Pearson correlation and estimated over 3000 replications with a sample size of 1000.

In total, twelve distinct combinations of the parameters α_0 and α_1 were selected in order to generate datasets which cover a wide range of different dependencies of X and Z. Furthermore, the following three group size conditions were chosen: equal group sizes [$P(X = 1) = 0.5$], unequal group sizes with the treatment group larger than the control group [$P(X = 1) = 0.8$], and unequal group sizes with the treatment group smaller than the control group [$P(X = 1) = 0.2$].

4.2.2 Outcome Model

While the treatment assignment was generated with respect to the mean-centered covariate Z^* with a unit variance, transformed covariates with expectations different from zero were used for the outcome model: $Z = \mu_Z + \sigma_Z \cdot Z^*$. Two values for the expectation $E(Z) = \mu_Z$ were selected and incorporated in both parts of the simulation study.[2] The variance of the covariate, i. e., $Var(Z) = \sigma^2_Z$, was fixed at the value of one for all conditions.

The datasets were generated in such a way that the covariate-treatment regression $E(Y|X,Z)$ was always Z-conditionally unbiased. The following *linear parameterization* of the regression of the outcome

[1] The interdependence of Nagelkerke's $R^2_{X|Z}$, the Pearson correlation coefficient $Cor(X,Z)$, and the group size $P(X = 1)$ as a functions of α_0 and α_1 is also visualized in the additional Figure 1 on page 12 of the digital appendix.

[2] In simulation study I a value of $\mu_Z = 10$ was used for data generation, and in study II datasets were generated with $\mu_Z = 5$. Within each part of the Monte Carlo study the expectation of the covariate $E(Z)$ was not manipulated as an additional factor of the simulation design.

```
# Covariate

z_star        <- rnorm(n,0,1)
z             <- mean.tau_z + sqrt(var.tau_z) * z_star

# Assignment Model

pscore        <- 1 -  1 / (1 + exp(alpha0 + alpha1 * z_star))
x             <- 0.0 + (runif(n) <= pscore)

# Outcome Model

eps_tau_0     <- rnorm(n, mean=0, sd=sqrt(var.eps_tau_0))
eps_delta_10  <- rnorm(n, mean=0, sd=sqrt(var.eps_delta_10))
zeta          <- rnorm(n, mean=0, sd=sqrt(var.zeta))

tau_0         <- ga00 + ga01 * z + eps_tau_0
delta_10      <- ga10 + ga11 * z + eps_delta_10

y             <- tau_0 + delta_10 * x + zeta
```

Listing 4.1: R syntax for the data generation

variable Y on treatment variable X (i. e., for the covariate-treatment regression) with Z as univariate numerical covariate was selected as the *functional form*:

$$\begin{aligned} E(Y|X,Z) &= E(\tau_0|Z) + E(\delta_{10}|Z) \cdot X \\ &= \left(\gamma_{00} + \gamma_{01} \cdot Z\right) + \left(\gamma_{10} + \gamma_{11} \cdot Z\right) \cdot X. \end{aligned} \tag{4.2}$$

As discussed as one of the implications of the theory of stochastic causality in section 3.1, we can differentiate $\zeta \equiv Y - E(Y|X,Z)$, $\varepsilon_{X=1} \equiv \tau_0 - E(\tau_0|Z)$ and $\varepsilon_{\delta_{10}} \equiv \delta_{10} - E(\delta_{10}|Z)$ as residual terms for a dichotomous treatment variable X:

$$\begin{aligned} Y = \;& E(Y|X,Z) + \zeta \\ = \;& \left(E(\tau_0|Z) + \varepsilon_{X=0}\right) \cdot I_{X=0} + \left(E(\tau_1|Z) + \varepsilon_{X=1}\right) \cdot I_{X=1} + \zeta \\ = \;& \left(E(\tau_0|Z) + \varepsilon_{X=0}\right) + \left(E(\delta_{10}|Z) + (\varepsilon_{X=1} - \varepsilon_{X=0})\right) \cdot X + \zeta \\ = \;& \left(E(\tau_0|Z) + \varepsilon_{X=0}\right) + \left(E(\delta_{10}|Z) + \varepsilon_{\delta_{10}}\right) \cdot X + \zeta. \end{aligned} \tag{4.3}$$

The three residuals $\varepsilon_{X=0}$, $\varepsilon_{\delta_{10}}$ and ζ (respective their variances) cannot be identified in empirical applications with the methods discussed in this thesis. Nevertheless, the individual total effect $\delta_{10} = \tau_1 - \tau_0$ can

alternatively be expressed as a regression on the covariate, i. e., for the linear parameterization presented in Equation (4.2) as

$$\begin{aligned} E(\delta_{10}|Z) = &\quad E(\tau_1|Z) - E(\tau_0|Z) \\ = &\quad \gamma_{10} + \gamma_{11} \cdot Z \text{ , with} \\ \varepsilon_{\delta_{10}} \equiv &\quad \delta_{10} - E(\delta_{10}|Z) \text{ , and} \\ Var(\delta_{10}) = &\quad Var(\gamma_{10} + \gamma_{11} \cdot Z + \varepsilon_{\delta_{10}}) \\ = &\quad \gamma_{11}^2 Var(Z) + Var(\varepsilon_{\delta_{10}}). \end{aligned} \tag{4.4}$$

According to the parameterization described above, the complete generation for the outcome variable Y can be summarized as:

$$Y = \left(\gamma_{00} + \gamma_{01} \cdot Z + \varepsilon_{X=0}\right) + \left(\gamma_{10} + \gamma_{11} \cdot Z + \varepsilon_{\delta_{10}}\right) \cdot X + \zeta. \tag{4.5}$$

Technically, the outcome variable was generated by segmenting Equation (4.5) into small parts: In line 4 of Listing 4.1 the covariate Z is generated. In lines 13, 14, and 15 the values of the three residual variables are drawn from standard normal distributions. The values of the true outcome variable in the control condition τ_0 and the value of the individual total effect variable δ_{10} are drawn in lines 17 and 18. Finally, the outcome model is completed in line 20. In other words, the covariate-treatment regression $E(Y|X,Z)$ was generated as a linear moderated regression with heteroskedastic errors.[3]

Table 4.3: Regression coefficients and effect sizes used for data generation in simulations I and II

Simulation Study	Parameter	Selected Values for the Data Generation
I	γ_{01}	1, 5
	γ_{11}	0.5, 1, 2.5, 5, 7.5, 10
	d	0
II	γ_{01}	1, 5
	γ_{11}	1, 2.5, 5
	d	0, 0.2, 0.5, 0.8

For all conditions of simulation study I, data were generated with a true population value of zero for the average total effect, i. e., $ATE_{10} = E(\gamma_{10} + \gamma_{11} \cdot Z) = 0$ ($d = 0$). Therefore, the appropriate regression coefficient γ_{10} was computed as a function of γ_{11} and $E(Z)$, i. e., $\gamma_{10} = -\gamma_{11} E(Z)$. For simulation study II, the true average total effect was generated with different *effect sizes* based on Cohen's definition. Four different values for the effect size d were chosen for this part of the simulation study. In terms of d, a

[3] The individual residual variances are equal to $Var(\zeta) + Var(\varepsilon_{X=0})$ for all individuals assigned to treatment group $X = 0$ and $Var(\zeta) + Var(\varepsilon_{X=0}) + Var(\varepsilon_{\delta_{10}})$ for units assigned to $X = 1$. All residual terms are generated independently and are therefore assumed to be uncorrelated.

small effect ($d = 0.2$), a medium effect ($d = 0.5$) and a large effect ($d = 0.8$) were used for comparing the statistical power of the different implementations of generalized analysis of covariance. Furthermore, to investigate the small sample performance of the adjustment procedures, a condition with no true average total effect ($d = 0$) was also included in simulation study II. The different levels of the regression coefficients and the different effect sizes used for the data generation are summarized in Table 4.3.

To manipulate the factor *heterogeneity of residual variances*, the variances of the two normally distributed random variables $\varepsilon_{\delta_{10}}$ and $\varepsilon_{X=0}$ were varied with the values given in table Table 4.4. In all conditions of the Monte Carlo simulation, a constant value of 0.5 was used for the residual variance $Var(\zeta)$ [see Equation (4.3) for the exact meanings of these terms].

Table 4.4: Residual variances used for data generation in simulations I and II

Simulation Study	Residual Variance	Selected Values for the Data Generation
I	$Var(\varepsilon_{\delta_{10}})$	0.5, 1, 2.5, 5
	$Var(\varepsilon_{X=0})$	0.5, 1, 2.5, 5
	$Var(\zeta) = 0.5$	
II	$Var(\varepsilon_{\delta_{10}})$	0.5, 2.5, 5
	$Var(\varepsilon_{X=0})$	0.5
	$Var(\zeta) = 0.5$	

This method of data generation for the outcome model resulted in a larger variance of the outcome variable Y in the treatment group (indicated by $X = 1$) than in the control group ($X = 0$) across all generated datasets. This is notable, as we hereby introduce the important distinction between the two conditions with unequal treatment probabilities [$P(X = 1) = 0.2$ versus $P(X = 1) = 0.8$].

Table 4.5 summarizes all parameters used for data generation. Whenever the factor *heterogeneity of between-group residual variances* is mentioned in the result presentation, it refers to the ratio of $Var(\varepsilon_{X=0})$ to $Var(\varepsilon_{\delta_{10}})$ as used for the data generation in the outcome model. Furthermore, we will use the phrase *amount of confounding* to distinguish the results obtained under the two different levels of the regression parameter γ_{01}. *Equal group sizes* [$P(X = 1) = 0.5$] and *unequal group sizes* [$P(X = 1) = 0.2$ or $P(X = 1) = 0.8$] are obtained from the coefficients α_0 and α_1 of the assignment model (see Table 4.2). Finally, note that the selection of the two α–parameters also determines the factor *dependency of X and Z*, labeled as $R^2_{Y|Z}$ within the result sections.

4.3 Design of the Simulation Studies

The simulation study was conducted in a fully crossed design, with $N_{\text{Rep}} = 1000$ replications of each combination of the varied independent parameters. For the first part of the simulation study (bias of the *ATE*–estimators, standard error bias of the *ATE*–estimators, and empirical type-I-error rate for the test of the

Table 4.5: Summary of the parameters used for data generation in simulations I and II

Sample Size	N	Number of observations
Regression Coefficients	γ_{00} / β_{00}	Intercept of the covariate-treatment regression in the control group
	γ_{01} / β_{01}	Slope of the covariate-treatment regression in the control group
	$\gamma_{10} = \beta_{10} - \beta_{00}$	Main effect (average total effect when no covariate-treatment interaction is present)
	$\gamma_{11} = \beta_{11} - \beta_{01}$	Covariate-treatment interaction (difference between the slopes of the group-specific covariate-treatment regressions)
	$\beta_{10} = \gamma_{10} + \gamma_{00}$	Intercept of the covariate-treatment regression in the treatment group
	$\beta_{11} = \gamma_{11} + \gamma_{01}$	Slope of the covariate-treatment regression in the treatment group
	α_0	Intercept for the assignment model
	α_1	Slope for the assignment model
Expectations	$E(Z) = \mu_Z$	Expectation of the covariate
Variances	$Var(Z) = \sigma^2_Z = 1$	Variance of the covariate
Residual Variances	$Var(\varepsilon_{X=0})$	Residual variance for the covariate-treatment regression in the control group
	$Var(\varepsilon_{X=1})$	Residual variance for the covariate-treatment regression in the treatment group
	$Var(\varepsilon_{\delta_{10}})$	Residual variance for the individual total effect
	$Var(\zeta)$	Additional residual variance of the outcome model (uncorrelated with the residual variance for the covariate-treatment regression in the control group)
Covariances	$Cov(\varepsilon_{X=0}, \varepsilon_{X=1}) = 0$	Covariance of the residuals for the treatment group-specific regressions of Y on Z
	$Cov(\varepsilon_{X=0}, \varepsilon_{\delta_{10}}) = 0$	Covariance of the residual for the regression of Y on Z in the control group with the residual of the regression δ_{10} on Z

hypothesis $ATE = 0$), 9216 cells were generated by combining the factor *sample size* (4 levels, see Table 4.1), the *dependency of X and Z*, the *group size* (12 combinations of α_0 and α_1 for the assignment model, see Table 4.2), the *regression coefficients* of the outcome model ($2 \cdot 6$ combinations of γ_{01} and γ_{11}, see Table 4.3), and the residual variances of the outcome model [$4 \cdot 4$ different pairs of $Var(\varepsilon_{X=0})$ and $Var(\varepsilon_{\delta_{10}})$, see Table 4.4].

The statistical power and the sample size requirements of the final models (simulation II) were studied in 8640 different cells. According to Table 4.1, the *sample size* was manipulated with 10 different levels, crossed with 12 combinations of α_0 and α_1 for the assignment model. Furthermore, the outcome model was generated for $2 \cdot 3$ different combinations of γ_{01} and γ_{11} (see Table 4.3) and 3 different residual variances $Var(\varepsilon_{\delta_{10}})$ [see Table 4.4]. Finally, 4 different values of the effect size d were used (see Table 4.3).

For each cell in the first and second part of the Monte Carlo simulation the following implementations of generalized analysis of covariance were applied to estimate the average total effects and to test the hypothesis $ATE = 0$ in the N_{Rep} simulated datasets:

- Two tests of the hypothesis of no average total effect implemented with the general linear hypothesis, either based on the estimated empirical mean of the covariate in the linear hypothesis, or with the true expectation of the covariate (see subsection 3.2.3 for details)
- The hypothesis $ATE = 0$ tested with the help of the general linear hypothesis and based on the mean-centering procedure, but with heteroskedasticity–adjusted variance-covariance matrices (`HC3` and `HC4` correction, see page 58 in subsection 3.2.2.1 for details)
- A test statistic for the estimated average total effect obtained as regression estimate and performed with the corresponding adjusted standard errors given by Schafer and Kang (2008) [see page 58 in subsection 3.2.2.1 for details]
- The application of the predictive simulation approach suggested by Gelman and Hill (2007) [see page 61 in subsection 3.2.3 for details]

Furthermore, for each generated dataset the hypothesis $ATE = 0$ was tested with the Wald–test of the non-linear constraint based on the following structural equation models:

- The *simple multi-group model* with fixed group size, where the mean of the treatment variable is assumed to be a known number (either from the data generation as the true population value or as the estimated empirical mean of the treatment variable X, see subsection 3.3.3.1 for details)
- The *elaborated multi-group model* as an extension of the simple multi-group model, where the group size is incorporated as an additional estimated model parameter with the `KNOWNCLASS`–option of `Mplus` (see subsection 3.3.3.2 for details)
- The *approximated multi-group model* with augmented variance-covariance matrix of parameter estimates (see subsection 3.3.3.3 for details)
- The *simple single group model (with interaction)* [see section 3.3.4.1 for details]
- The *elaborated single group model*, where the implied variance structure is modeled with the help of the random slope approach (see section 3.3.4.2 for details)

The following methods were applied only to the generated datasets of the first part of the simulation study to save computational time: The simple multi-group model with fixed group size utilizing the true group size $P(X = 1)$ known from the data generation and the general linear hypothesis / the moderated regression approach with a mean-centered covariate with and without heteroskedasticity–adjusted variance-covariance matrices.

For both parts of the simulation study `R` was used for data generation, data management and regression modeling (R Development Core Team, 2008). All structural equation models were estimated with `Mplus` Version 5.0 (L. K. Muthén & Muthén, 1998 - 2007).

4.4 Dependent Measures

4.4.1 Bias of the *ATE*–Estimator

The mean bias of the average total effect estimator, defined as the difference between the mean of the estimated average total effect over all replications, i. e., $\overline{\widehat{ATE}_{10}}$, and the true population value of the average total effect, i. e., ATE_{10}, was computed for each cell of the simulation study:

$$B(\widehat{ATE}_{10}) = \overline{\widehat{ATE}_{10}} - ATE_{10}. \tag{4.6}$$

The true parameters for the data generation, as described in detail in the last subsections, were selected to yield true average total effects of zero for all conditions of simulation study I. Therefore, it was not necessary to standardize this absolute bias.

4.4.2 Relative Bias of the Standard Error of the *ATE*–Estimator

The accuracy of the estimated standard error of the *ATE*–estimator was evaluated as the mean *relative bias* (see, for example, Boomsma & Hoogland, 2001):

$$RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right] = \frac{\overline{\widehat{\text{S.E.}}}\left(\widehat{ATE}_{10}\right) - SD\left(\widehat{ATE}_{10}\right)}{SD\left(\widehat{ATE}_{10}\right)}. \tag{4.7}$$

This relative bias can be interpreted as the percentage of bias in the standard error relative to the true variability of the estimator. This measure is independent of the true parameter value and can be compared for datasets with different sample sizes. The mean of the estimated standard errors, $\overline{\widehat{\text{S.E.}}}\left(\widehat{ATE}_{10}\right)$, and the standard deviation of the estimates of the average total effect, $SD\left(\widehat{ATE}_{10}\right)$, are calculated over the number of replications (N_{Rep} = 1000) for each cell of the simulation study (Lei & Lomax, 2005; D. P. MacKinnon, Lockwood, Hoffmann, West, & Sheets, 2002; Nevitt & Hancock, 2004).[4]

[4] Note that sometimes criteria are reported for the judgement of ignorable relative standard error biases (e. g., Boomsma & Hoogland, 2001, used $|RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]| < 0.05$ and Nevitt & Hancock, 2004, used $|RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]| < 0.1$). We do not apply these criteria to the results of the simulation study because according to our knowledge the cutoff values were developed without a clear statistical rationale. Nevertheless, we will highlight values of the relative standard error bias $|RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]| > 0.05$ with bold numerics, when comparing the results of simulation study I with the biases reported by Flory (2008).

4.4.3 Type-I-Error Rate for Testing $H_0 : \boldsymbol{ATE} = 0$

We computed the *rejection frequency* (h_{RF}) to compare the empirical type-I-error rates and the statistical power of different implementations of generalized analysis for a test of the hypothesis $ATE = 0$ (within each cell of the simulation study and based on the commonly used α–level of 0.05).

For a given number of replications N_{Rep}, the expected rejection frequency h_{RF} can be described as a binominal distributed random variable, i. e., $h_{\text{RF}} \sim Bin(N_{\text{Rep}}, \alpha)$. A 95% and a 99% confidence interval can be constructed around the nominal α – level with the quantile function (see Boomsma & Hoogland, 2001, for details).[5] In all plots of the observed empirical distribution of the rejection frequencies presented in the result sections, red lines (99%) and gray lines (95%) refer to these confidence intervals.

4.4.4 Statistical Power for Testing $\boldsymbol{ATE} = 0$

The statistical power of the studied test statistics, and especially of the Wald–tests obtained from nonlinear constraints of model parameters within the framework of structural equation modeling, is influenced by a multitude of factors (see, for example, L. K. Muthén & Muthén, 2002). Nevertheless, because of the simplicity of the analyzed regression model (a manifest regression model without any latent variables, and only one covariate and a simple comparison of two groups) the power analysis conducted in the second part of the simulation study differs from typical power analyses, as mainly discussed in the literature of structural equation modeling (see, e. g., Hancock & Mueller, 2006). The methodology of structural equation modeling is merely a device to estimate simple regression models and to test hypotheses about nonlinear functions of model parameters, under the assumption of a multivariate joint distribution of Y and Z or of Y, Z and X. Therefore, the usual complications of hypothesis testing and power problems in structural equation models with latent variables do not apply to the model comparison we provide with respect to the research question described in subsection 3.4.7. Furthermore, the goal of the power analysis performed for this thesis is not to determine how large a sample must be to achieve a reasonable chance of rejecting a specified model, but to compare the statistical power and the small sample behavior of the discussed implementations of generalized analysis of covariance for designs with stochastic regressors and covariate-treatment interactions.

Within the tradition of structural equation modeling, elegant methods of power analysis without excessive simulations have been developed, for instance, by Satorra and Saris (1985). We did not make use of these alternative strategies for two reasons: Firstly, because these procedures assume a correctly specified model. As described in subsection 3.3.3, some of the considered structural equation models are misspeci-

[5] See also the additional Figure 2 on page 13 of the digital appendix.

fied at least with respect to the implied variance structure (i. e., second order misspecifications, see Long & Trivedi, 1992). Therefore, this underlying assumption of the Satorra–Saris approach appears to be an untenable prerequisite. Secondly, we are interested in a comparison of structural equation models and methods based on (adjusted) ordinary least-squares estimates. In order to achieve this goal, we will use a more general "brute force" strategy and compare the empirically observed rejection frequencies between the different implementations of generalized analysis of covariance for datasets generated with a true average total effect different from zero.

4.4.5 Further Measures

To compare the different implementations of generalized analysis of covariance in detail, we also analyze the *mean squared errors* as measures of the relative efficiency of the *ATE*–estimation, and as a measure of the relative efficiency of the estimation of the corresponding standard errors, i. e.,

$$MSE\left[\widehat{ATE}_{10}\right] = \frac{\sum_i^{N_{\text{Rep}}} (\widehat{ATE}_{10i} - ATE_{10})^2}{N_{\text{Rep}}} \text{ , and} \tag{4.8}$$

$$MSE\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right] = \frac{\sum_i^{N_{\text{Rep}}} \left[\widehat{\text{S.E.}}_i\left(\widehat{ATE}_{10}\right) - SD\left(\widehat{ATE}_{10}\right)\right]^2}{N_{\text{Rep}}}. \tag{4.9}$$

Furthermore, we will report *convergence rates* for the different structural equation models. The convergence rate is defined as the number of converged runs divided by the number of replications per cell of the simulation design and will be reported in percentage. Empirical type-I-error rates and rejection frequencies for tests of the hypothesis $ATE = 0$ (used to report the adherence to the nominal α–level and to compare the statistical power of the different methods) are computed always based on the converged replications within each cell of the simulation design.

Finally, *empirical covariances of parameter estimates* are computed for the multi-group model as a measure of the true dependency of parameter estimates.

4.5 Results for the General Linear Model

The results of the Monte Carlo simulation will be presented in the following four sections. This first subsection deals with the results obtained for the ordinary least-squares estimated covariate-treatment regression: Different test statistics for the hypothesis $ATE = 0$ that are based on the unadjusted general linear model (i. e., mean-centering and the general linear hypothesis), the test statistics obtained from heteroscedasticity adjusted ordinary least-squares regression, the test statistic based on the adjusted standard errors for the average total effect estimated as a regression estimate, and the test statistic for the average total effect based

on predictive simulations. For each of the adjustment procedures (see also subsection 4.3) and separately for all cells of the simulation study, we computed the dependent measures described in section 4.4.

In this chapter, nevertheless, we will report on only a selection of the most important results, which are necessary to answer the specific research questions given in section 3.4. Supplementary plots generated from data of part I and part II of the simulation study are provided in the digital appendix and as a digital supplement on DVD (see page 245 in the appendix for the content and for the structure of these additional materials).

4.5.1 General Linear Hypothesis and Mean-Centering

According to the research question formulated in subsection 3.4.1, the results of the general linear model are reported here to demonstrate the expected bias of the standard errors of the *ATE*–estimator caused by the stochasticity of the covariates as well as to provide empirical evidence that mean-centering of covariates does not change the statistical properties of the ordinary least-squares estimated average total effect estimator (e. g., the standard error). Furthermore, we chose the robustness of the general linear model to heterogeneity of residual variance for unequal group sizes as one of the central themes for the simulation study I (see the reasearch question in subsection 3.4.2). Hence, it will be interesting to see if the two violations of the assumptions of the general linear model (i. e., the violation of the fixed-X assumption and the violation of the heteroskedasticity assumption) can be distinguished by contrasting the effects of different parameter constellations of the Monte Carlo simulation.

First of all, note that we obtained the same results for the mean-centering approach and for the procedures based on the general linear hypothesis with respect to the absolute bias of the *ATE*–estimator [$B(\widehat{ATE}_{10})$, as defined in Equation (4.6)] and moreover with respect to the relative bias of the standard error of the *ATE*–estimator [$RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$, as defined in Equation (4.7)] and accordingly also with respect to the empirical type-I-error rates (i. e., the observed rejection frequencies, h_{RF}) for tests of the hypothesis $ATE_{10} = 0$.[6]

Absolute Bias of the *ATE*–Estimator The *ATE*–estimator of the GLH / mean-centering approach was found to be unbiased for all studied conditions of simulation study I.[7] Figure 4.1 presents the $B(\widehat{ATE}_{10})$ for the GLH / mean-centering approach based on the estimated mean of the covariate on the y–axis, compared to

[6] Negligible differences between the mean-centering approach and the general linear hypothesis — if one method was implemented based on the GLM implementation of the general linear model in R, and the other method was implemented based on the LM implementation of the general linear model in R— are reported as additional Figure 3 on page 14, Figure 4 on page 15, Figure 5 on page 16, Figure 6 on page 17, Figure 7 on page 18 and Figure 8 on page 19 of the digital appendix. All results reported in this result section are based on the LM implementation of R (see Chambers & Hastie, 1991).

[7] The absolute biases of the *ATE*–estimator obtained for the GLH / mean-centering approach are given in the additional Table 1 on page 121, Table 2 on page 122 and Table 3 on page 123 of the digital appendix. Furthermore, the additional Table 4 on page 124 of the digital appendix presents the $B(\widehat{ATE}_{10})$ averaged for all conditions of the simulation study I, grouped by different sample sizes. Apart from random fluctuations, the absolute biases of the average total effect estimator decrease as expected for increasing sample sizes.

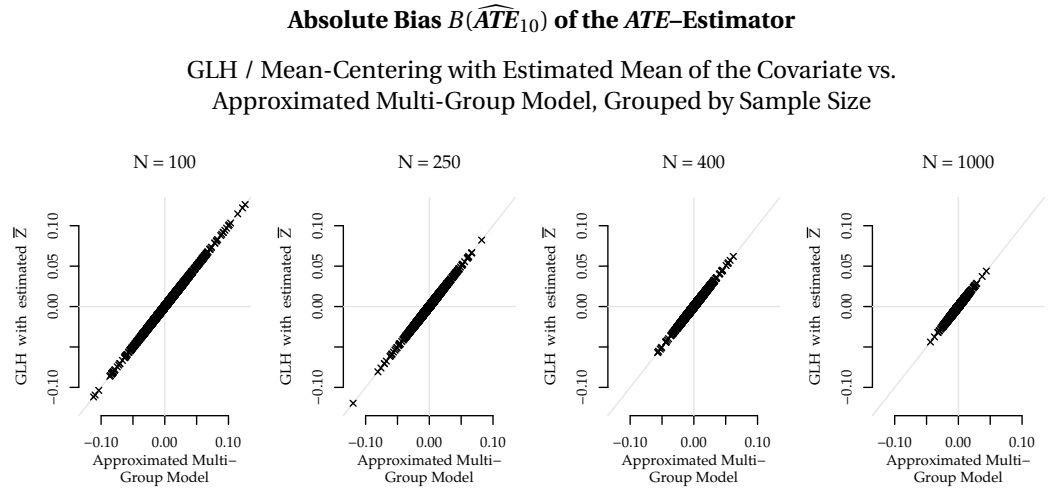


Figure 4.1: Absolute bias of *ATE*–estimator: Scatter plots for a comparison of the GLH / mean-centering approach (estimated mean of the covariate) and the approximated multi-group model, grouped by sample size N

the $B(\widehat{ATE}_{10})$ for the approximated multi-group structural equation model on the x–axis. The scatter plots are grouped by different values of the sample size N, that means that each sample size condition of the simulation study I ($N = 100$, $N = 250$, $N = 400$, and $N = 1000$) is plotted separately as a chart. Each symbol in the four charts represents the $B(\widehat{ATE}_{10})$ for one condition of the simulation study I, located in the chart according to the observed absolute biases of the two mentioned *ATE*–estimators. It is obvious that both methods are comparable with respect to the absolute bias of the average total effect estimator because all symbols are located perfectly on the diagonal line for each of the presented sample size conditions.

A comparison of the $B(\widehat{ATE}_{10})$ for two different sample sizes ($N = 100$ top, versus $N = 1000$, bottom) is summarized in Figure 4.2. This figure presents the biases for 288 conditions of simulation study I with $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 0.75$ for the GLH / mean-centering approach, condensed as *level plots*.[8] The $B(\widehat{ATE}_{10})$ is different from zero for conditions with small sample sizes (obvious from a comparison of the first and the third row in the upper part of the figure with the same two rows in the lower part of the figure), especially for unequal group sizes with a treatment probability of $P(X = 1) = 0.2$.

Replacing the estimated mean of the covariate, i. e., $\widehat{\mu}_Z$, with the expectation of the covariate, i. e., $E(Z)$, yields different biases for the *ATE*–estimator, especially for data generated with large interaction effects. This observation is presented as scatter plots in Figure 4.3.[9] Each of the six charts in this figure

[8] Quantitative results, for instance, the observed bias of the *ATE*–estimator, are encoded in these level plots by using a color gradient. The corresponding color key for mapping the printed colors to the observed values is given on the right side of each plot. Level plots are used often to present results because they are useful for judging patterns of variability, see Sarkar, 2008, for more information about level plots and their generation with R.

[9] Furthermore, the additional Figure 9 on page 20 of the digital appendix compares the observed distribution of $B(\widehat{ATE}_{10})$ based on the estimated mean of the covariate in the left column and based on the true expectation of the covariate in the right column,

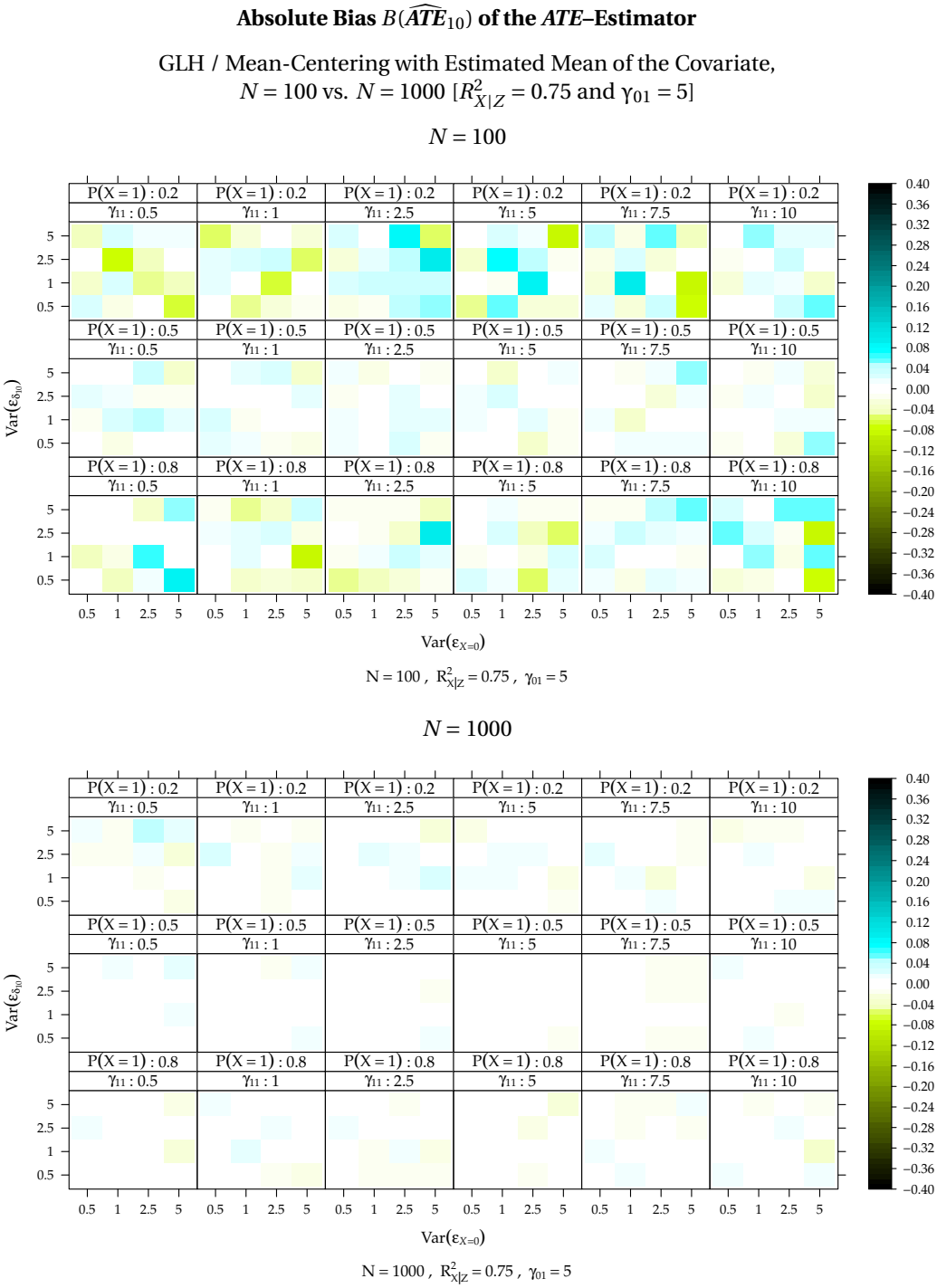


Figure 4.2: Absolute bias of *ATE*–estimator: Level plots for the GLH / mean-centering approach based on the estimated mean of the covariate [$N = 100$ vs. $N = 1000$; $R^2_{X|Z} = 0.75$, $\gamma_{01} = 5$ and $\gamma_{11} = 0.75$]

presents the $B(\widehat{ATE}_{10})$ for the GLH / mean-centering procedure with $\widehat{\mu}_Z$ on the x–axis, and for the GLH /

conditional on the value of the interaction parameter γ_{11} used for data generation (rows). Obviously, the differences between the to GLH / mean-centering procedures increase with the interaction term γ_{11}.

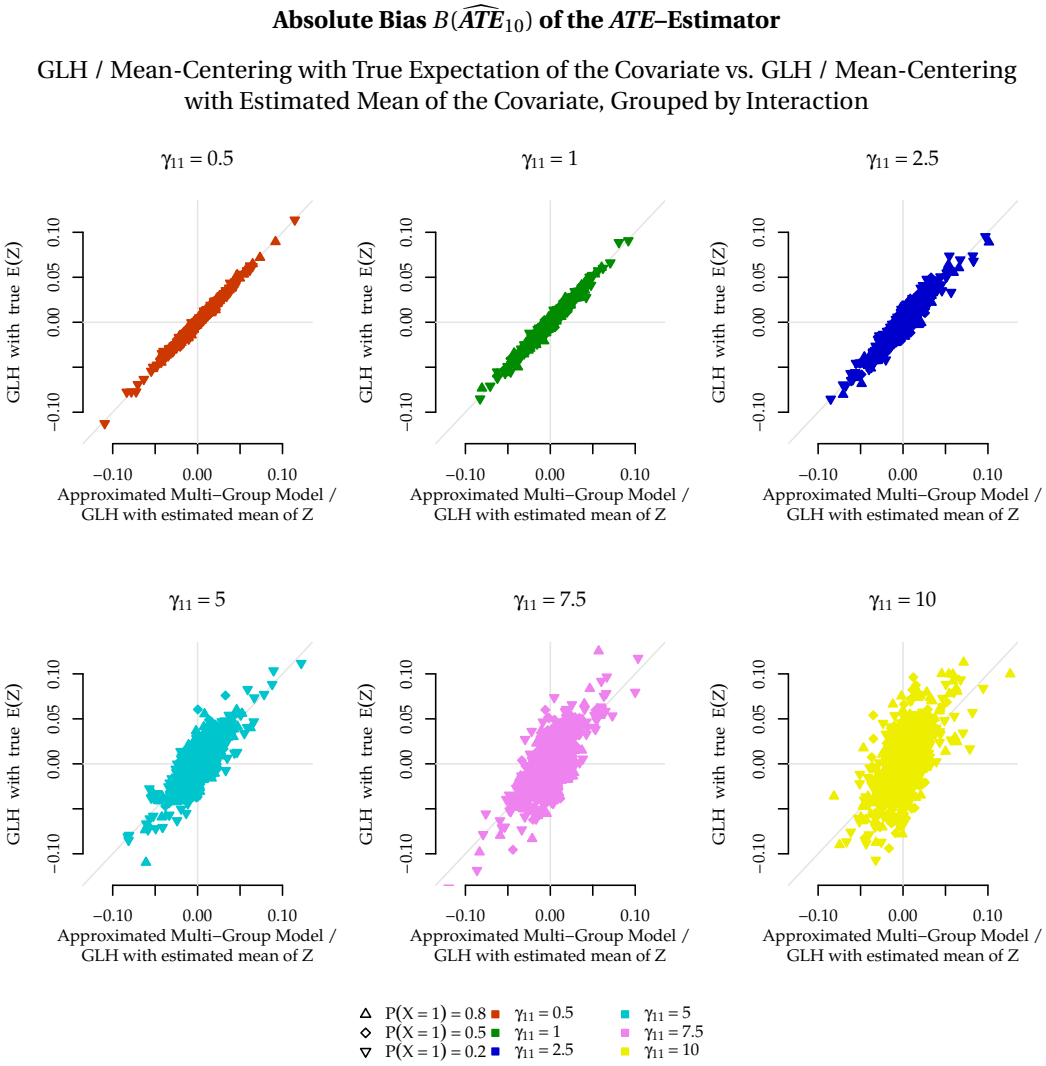
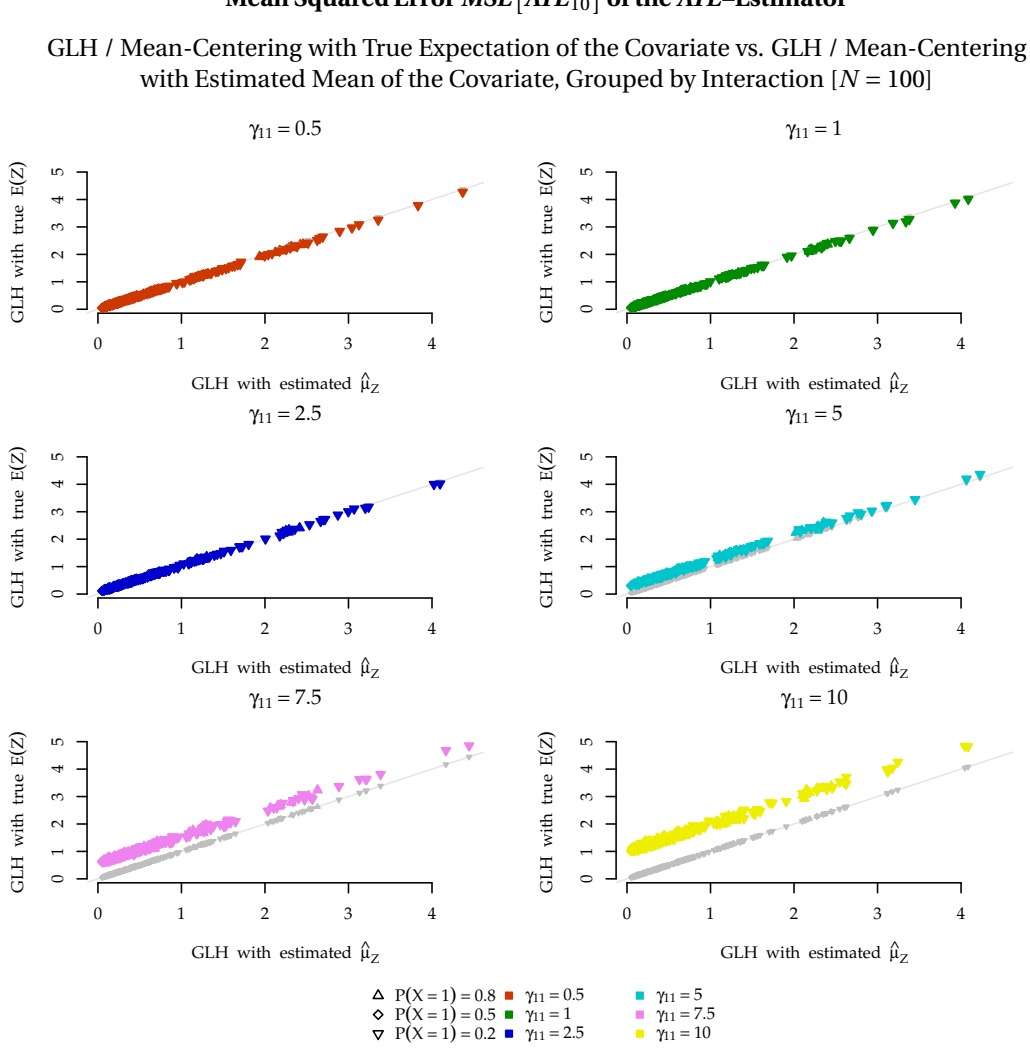


Figure 4.3: Absolute bias of the *ATE*–estimator: Scatter plots for the GLH / mean-centering approach (true expectation of the covariate vs. estimated mean of the covariate), grouped by interaction γ_{11}

mean-centering procedure with $E(Z)$ on the y–axis. The charts are grouped by the value of the interaction parameter γ_{11} and different symbols are used to distinguish the three different group size conditions. Although both estimates agree with each other for small values of γ_{11}, differences are observable for large interaction effects. Obviously, both procedures are more distinct for unequal group sizes [$P(X = 1) = 0.2$ and $P(X = 1) = 0.8$].

Mean Squared Error of the *ATE*–Estimator We now show that although the average total effect estimator is not biased for the GLH / mean-centering approach with $E(Z)$, the mean squared errors of the estimator are larger when the estimated mean of the covariate $\widehat{\mu}_Z$ is replaced by the true expectation of the covariate.



Note: Small gray symbols represent the results for the approximated multi-group model on the y–axis.

Figure 4.4: Mean squared error of the *ATE*–estimator: Scatter plots for the GLH / mean-centering approach (true expectation of the covariate) vs. GLH / mean-centering approach (estimated mean of the covariate), grouped by interaction γ_{11} [$N = 100$]

Figure 4.4 presents the $MSE[\widehat{ATE}_{10}]$ for a sample size of $N = 100$:[10] On the x–axis for the GLH / mean-centering procedure with $\hat{\mu}_Z$, and on the y–axis for the same method with $E(Z)$. As expected, the mean squared errors of the *ATE*–estimator are smaller for equal group sizes compared to unequal group sizes (the group size conditions are represented by different symbols in the figure). For large interaction effects, the $MSE[\widehat{ATE}_{10}]$'s differ remarkably between the two alternative implementations of generalized analysis of covariance (different values of γ_{11} are represented by different colors in the figure). The approach with

[10] The same structure is observed for $N = 1000$, see the additional Figure 10 on page 21 of the digital appendix.

$E(Z)$ yields consistently larger $MSE\left[\widehat{ATE}_{10}\right]$ for large interaction effects, i. e., this approach is less efficient when covariate-treatment interactions are present.[11]

Figure 4.5: Type-I-error rate: Scatter plots for the GLH / mean-centering approach (true expectation of the covariate vs. estimated mean of the covariate), grouped by $Var\left(\varepsilon_{\delta_{10}}\right)$ $[R^2_{X|Z} = 0.1]$

Type-I-Error Rate All datasets in the first part of the Monte Carlo study were generated with no (true) average total effect in the population. Therefore, the observed rejection frequencies (h_{RF}) are expected to correspond to the nominal α–level of 5 %. The h_{RF}'s are plotted in Figure 4.5 for all simulated conditions of part I of the simulation study with a dependency of X and Z equal to $R^2_{X|Z} = 0.1$. The four scatter plots are

[11] The interdependence of the $MSE\left[\widehat{ATE}_{10}\right]$ and γ_{11} is summarized as level plot in the additional Figure 11 on page 22 of the digital appendix, in which the $MSE\left[\widehat{ATE}_{10}\right]$'s for the GLH / mean centering approach with $\hat{\mu}_Z$ are plotted in the upper part, and the $MSE\left[\widehat{ATE}_{10}\right]$'s with $E(Z)$ are plotted in the lower part of the figure. Obviously, the $MSE\left[\widehat{ATE}_{10}\right]$'s of the GLH / mean-centering approach with $E(Z)$ increases with an increasing interaction parameter γ_{11}, especially for equal group sizes (as a comparison of the patterns in the second row between the upper and the lower part of the figure reveal).

grouped by different values of the residual variance $Var(\varepsilon_{\delta_{10}})$. The observed h_{RF}'s are further distinguished for the different levels of the interaction parameter used for data generation (γ_{11}, different colors indicating the amount of interaction), and for the three different treatment probabilities [$P(X = 1)$, different symbols indicating the group size], i. e., each dot represents the h_{RF} for one combination of the remaining parameters in simulation study I. The number of statistical significant tests of the GLH / mean-centering approach based on $E(Z)$ are shown on the *x*–axis (in percent), and the number of statistical significant tests based on $\widehat{\mu}_Z$ are shown on the *y*–axis (in percent).[12] For the interpretation of the results summarized in Figure 4.5 it is important to distinguish between the two conditions of equal and unequal group sizes: For equal group sizes [$P(X = 1) = 0.5$], the empircal type-I-error rates obtained for tests of the null hypothesis based on $\widehat{\mu}_Z$ are inflated if the interaction parameters γ_{11} are larger than one. For $\gamma_{11} \leq 1$ the procedures with $\widehat{\mu}_Z$ and $E(Z)$ yield comparable empirical type-I-error rates (the red and green symbols are around the diagonal).[13]

Replacing $\widehat{\mu}_Z$ with $E(Z)$ restricts the h_{RF} for equal group sizes [$P(X = 1) = 0.5$] within the confidence intervals. The empirical type-I-error rates are generally inflated for all simulated conditions with unequal group sizes [$P(X = 1) = 0.2$]. For the GLH / mean-centering procedure based on the estimated mean of the covariate, the amount of inflation depends on the interaction parameter γ_{11} (accordingly the colors in Figure 4.5 are ordered vertically), and for the theoretical alternative with the true expectation $E(Z)$ the observed inflation is connected to the amount of heterogeneity of residual variances [manipulated as residual variance $Var(\varepsilon_{\delta_{10}})$; visible as different widths of the scatter plots in the four charts in Figure 4.5]. For $P(X = 1) = 0.8$, the empirical type-I-error rates of the GLH / mean-centering procedure with the $E(Z)$ tend to be too small for homogenous residual variances, and the observed h_{RF}'s are clearly lower than the desired level for conditions with heterogeneous residual variances. For the approach based on $\widehat{\mu}_Z$, the empirical type-I-error rates are smaller than the nominal level for small interaction effects and are inflated for large interaction effects.

If the general linear hypothesis for the test of $ATE = 0$ is constructed with $E(Z)$, the obtained h_{RF}'s are valid for equal group sizes (see the second major row in the upper and lower part of Figure 4.6). The h_{RF}'s are either too high [for $P(X = 1) = 0.2$] or too low [for $P(X = 1) = 0.8$] for unequal group sizes due to the heterogeneous between-group residual variances. The dependency of the h_{RF}'s and the residual variances used for data generation can be observed as a diagonally increasing inflation of the empirical type-I-error rates for treatment groups smaller than control groups (the dominating pattern in the first row

[12] The additional lines in each plot mark the confidence intervals (see subsection 4.4.3), and the diagonal lines gives the identity.

[13] Similar plots for different values of $R^2_{X|Z}$ are given as additional figures, see Figure 12 on page 23, Figure 13 on page 24, and Figure 14 on page 25 of the digital appendix. A comparison of Figure 4.5 ($R^2_{X|Z} = 0.1$) with these additional figures, i. e., Figure 12 [$R^2_{X|Z} = 0.25$], Figure 13 [$R^2_{X|Z} = 0.5$], and Figure 14 [$R^2_{X|Z} = 0.75$], reveals that the differences between the GLH / mean-centering approach with $\widehat{\mu}_Z$ and $E(Z)$ are larger for small dependencies between X and Z. This means that the GLH / mean-centering procedure based on the estimated mean of the covariate is more robust to interaction effects if the covariate is strongly connected to the treatment assignment.

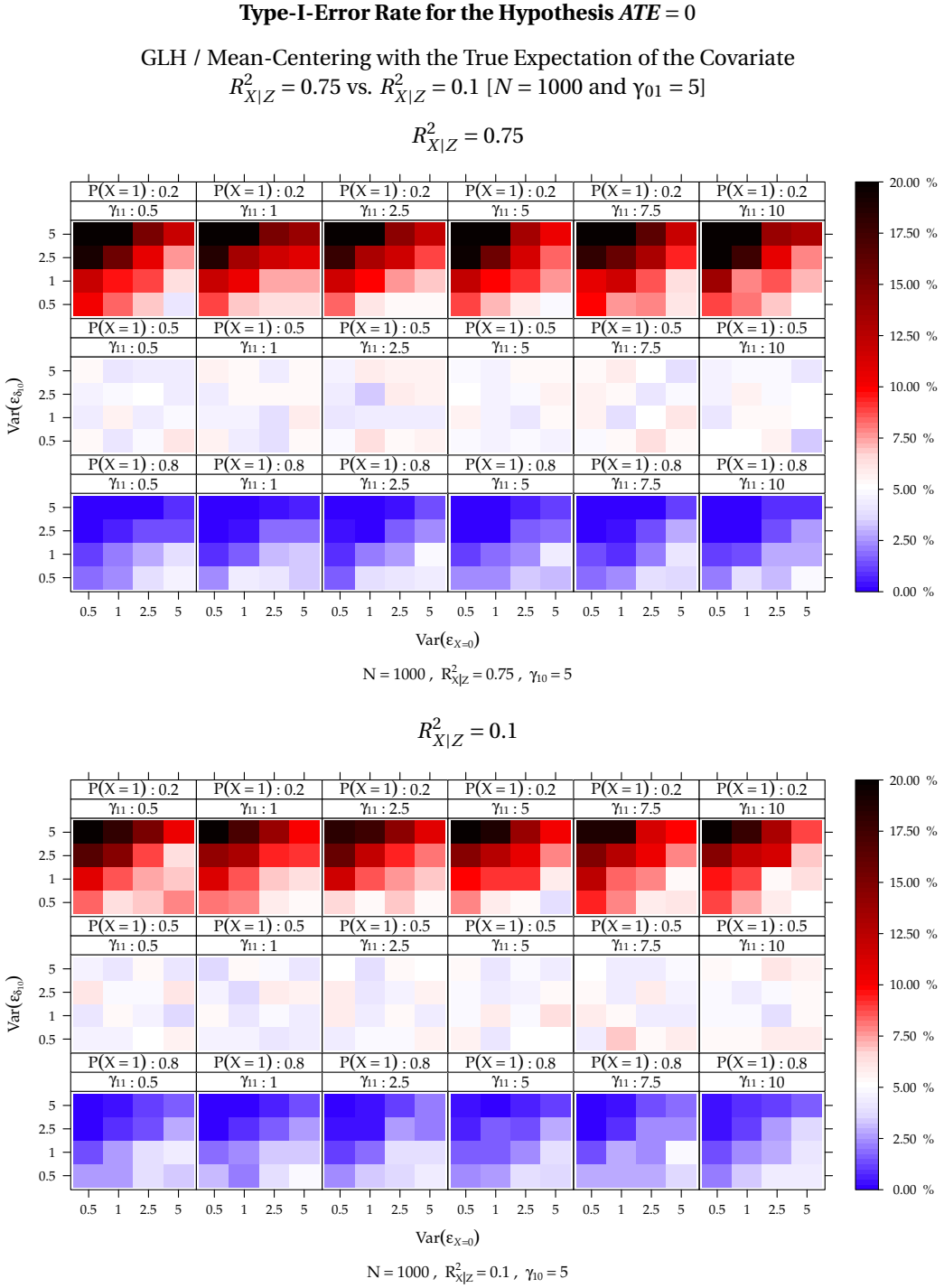


Figure 4.6: Type-I-error rate: Level plots for the GLH / mean-centering approach based on the true expectation of the covariate [$R^2_{X|Z} = 0.75$ vs. $R^2_{X|Z} = 0.1$; $N = 1000$ and $\gamma_{01} = 5$]

in the upper and the lower part of the figure). In the same way, the empirical type-I-error rates decrease systematically for treatment groups smaller than control groups. If the general linear hypothesis or the regression model for the mean-centering approach is constructed with $\widehat{\mu}_Z$, both the effect of heterogeneous

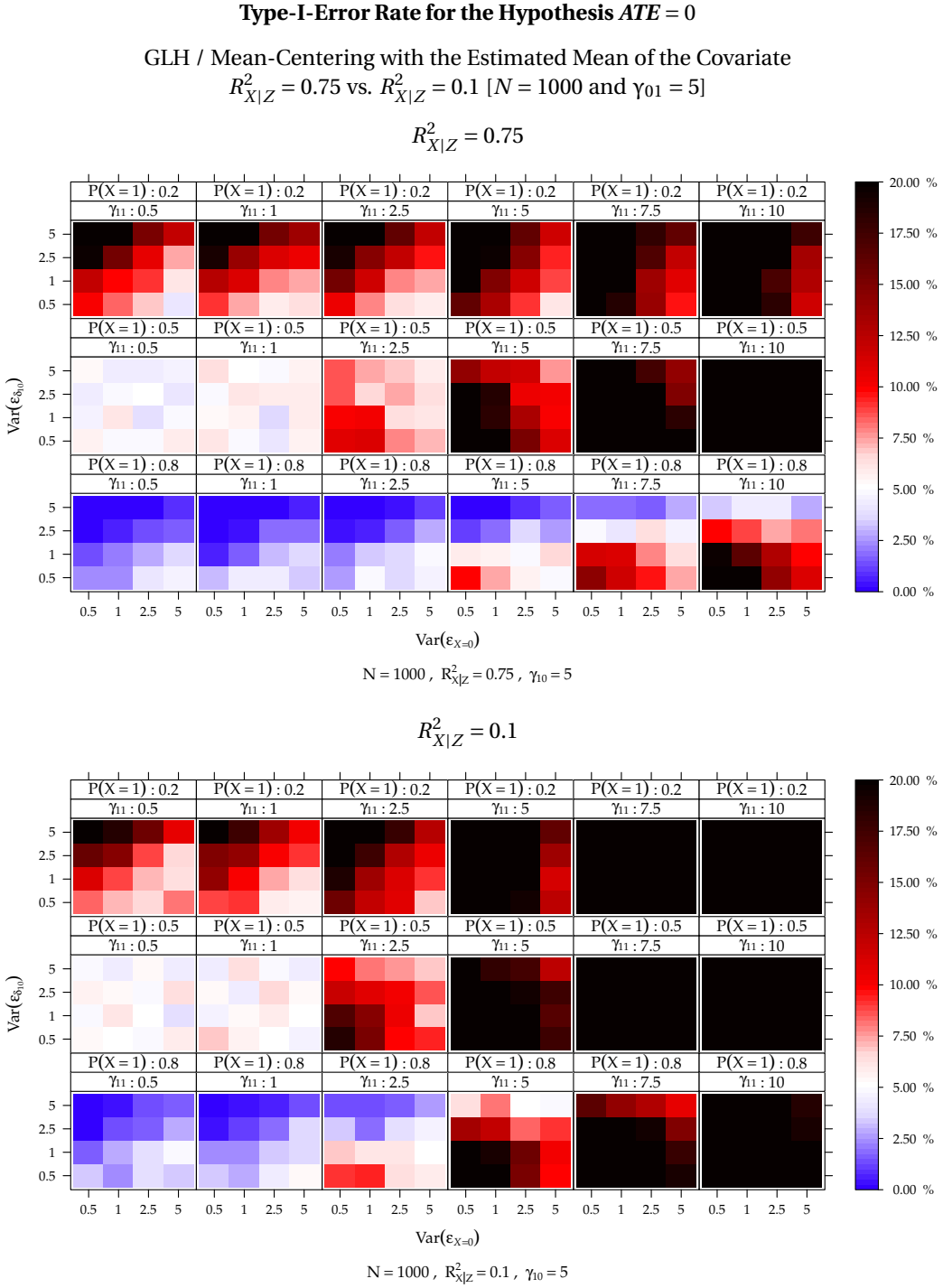


Figure 4.7: Type-I-error rate: Level plots for the GLH / mean-centering approach based on the estimated mean of the covariate [$R^2_{X|Z} = 0.75$ vs. $R^2_{X|Z} = 0.1$; $N = 1000$ and $\gamma_{01} = 5$]

residual variances and the consequences due to the nonlinearity of the hypothesis overlap with each other (see Figure 4.7). The observed pattern of the empirical type-I-error rates is inconspicuous for equal group sizes and small interaction terms ($\gamma_{11} = 0.5$ and $\gamma_{11} = 1$; see the first two columns of the second row in

the upper and lower part of Figure 4.7). For medium and large interaction terms ($\gamma_{11} > 1$), as predicted in section 3.2.5, the observed h_{RF}'s are inflated and increasing as a nonlinear (supposed to be quadratic) function of the interaction parameter γ_{11} (second rows in the figure).[14]

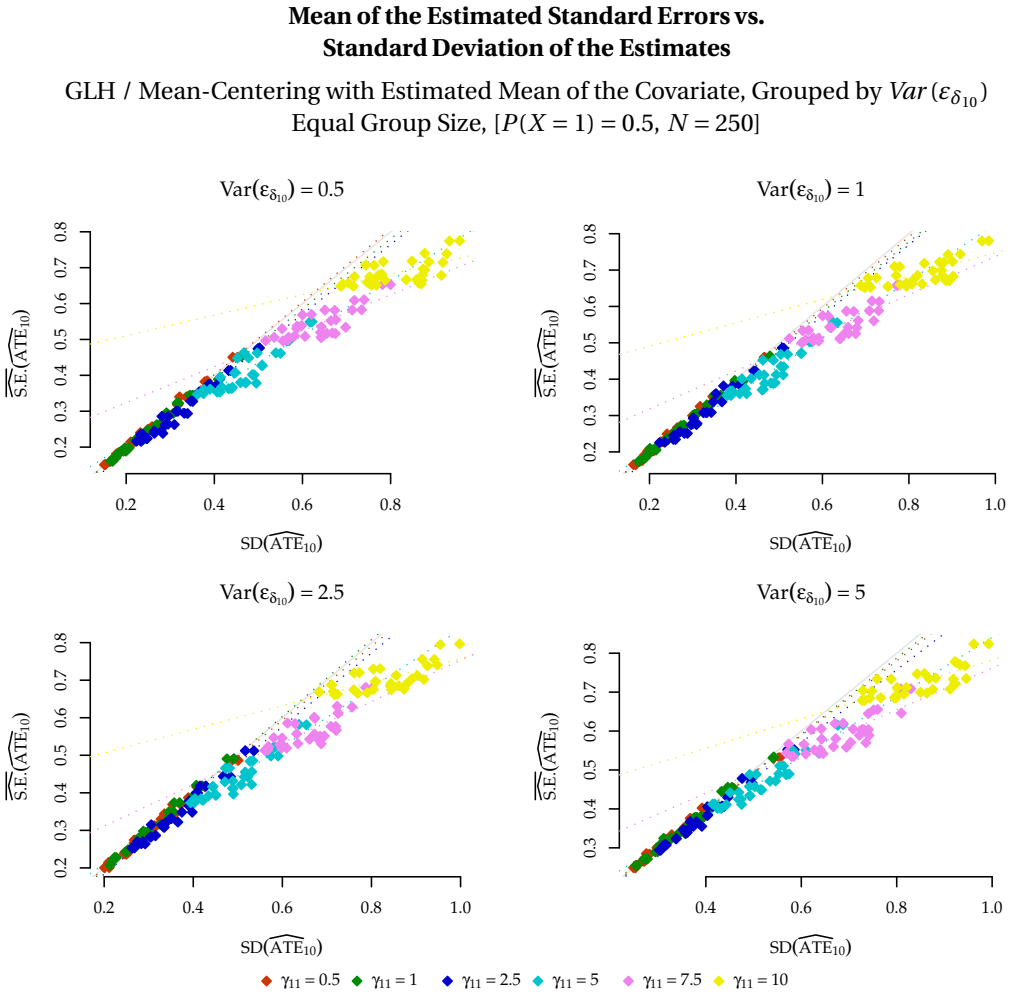


Figure 4.8: Mean of the estimated standard errors vs. standard deviation of the estimated average total effects, GLH / mean-centering (estimated mean of the covariate), grouped by $Var(\varepsilon_{\delta_{10}})$ [$N = 250$, $P(X = 1) = 0.5$]

Bias of the Standard Error of the *ATE*–Estimator We are now presenting the results for the relative bias of the *ATE*–estimators' standard error, $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$, in order to discover the reason for the inflated empirical type-I-error rates observed for the GLH / mean-centering approach based on the estimated mean of

[14] A comparison of the two conditions presented in the additional Figure 15 on page 26 of the digital appendix (strong dependency between X and Z in the upper part and low dependency between X and Z in the lower part) reveals that the inflation of the type-I-error rate depends on this part of the data generation, with more serious derivations of the observed h_{RF} for conditions where data were generated with a low level of dependency between X and Z. Finally, we conclude from the direct comparison of Figure 15 on page 26 of the digital appendix with Figure 4.7 that the amount of confounding (γ_{01}) is of minor importance only.

the covariates.[15] Figure 4.8 shows scatter plots of the two elements incorporated in the computation of the relative bias of the standard error of the *ATE*–estimator as defined in equation Equation (4.7): The mean of the estimated standard errors, i. e., $\overline{\widehat{\text{S.E.}}}\left(\widehat{ATE}_{10}\right)$, on the y–axis, and the observed standard deviation of the *ATE*–estimates for a given cell of the simulation study, i. e., $SD\left(\widehat{ATE}_{10}\right) = \sqrt{Var\left(\widehat{ATE}_{10}\right)}$, on the x–axis. The four charts in Figure 4.8 are grouped according to residual variances $Var\left(\varepsilon_{\delta_{10}}\right)$ used for the data generation. Furthermore, within each scatter plot of Figure 4.8, different colors represent different values of the interaction parameter γ_{11}. For a trustable estimator with unbiased standard errors it is expected that the mean of the estimated standard errors (y–axis) equals the standard deviation of the estimates of the average total effect (x–axis). This means that all symbols are expected to be on the diagonal line in the four scatter plots. This is clearly not the case, as the true variability of the average total effect estimator is underestimated for large interaction effects ($\gamma_{11} > 2.5$; all symbols are below the diagonal line). Nevertheless, both quantities increase as a function of the interaction (clearly visible as ordered colors). Hence, the results support the derivation that the standard errors and test statistics for procedures based on unadjusted ordinary least-squares estimates are incorrect as a consequence of the nonlinearity of the hypothesis if medium and large interactions are present (see section 3.2.5). For substantial interaction effects, the mean of the estimated standard errors for the GLH / mean-centering approach is generally smaller than the true variability of the average total effect estimates. Constructing a confidence interval based on the estimated standard errors for the GLH / mean-centering approach will result in heavily inflated type-I-error rates for unequal group sizes (exactly as we have described in the last subsection).[16]

The relative biases of the standard errors $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ in percent are printed in Figure 4.9 (grouped according to the different levels of heterogeneity of the between-group residual variances in columns and the amount of interaction used for data generation in rows). The figure shows the empirical distributions of the relative standard error biases of the average total effect estimator, approximated as histograms for a treatment probability $P(X = 1) = 0.8$.[17] These distributions of the standard error's relative bias are clearly bimodal for large interaction effects and strong heterogeneity of between-group residual variances (see, for example, the sixth row and the fourth column in Figure 4.9). We therefore disentangled these bimodal distributions by conditioning the results on the value of the dependency between X on Z (i. e., conditioning on $R^2_{X|Z}$) in Figure 4.10. It is obvious that the bias of the *ATE*–estimator's standard error depends on the strength of X and Z's association in the assignment model for large interaction effects. This finding clearly corrob-

[15] Note that the standard error for the average total effect estimator is the square root of the unconditional variance of the *ATE*–estimator as discussed in section 3.2.5.

[16] Note that for conditions with unequal group sizes, the two measures of the variability of the *ATE*–estimator differ even more. Detailed plots for these conditions are printed in the digital appendix, see the additional Figure 16 on page 27 for $P(X = 1) = 0.2$ and the additional Figure 17 on page 28 for $P(X = 1) = 0.8$.

[17] See the additional Figure 18 on page 29 and the additional Figure 19 on page 30 of the digital appendix for the distributions of the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for equal group sizes as well as for unequal group sizes with $P(X = 1) = 0.2$.

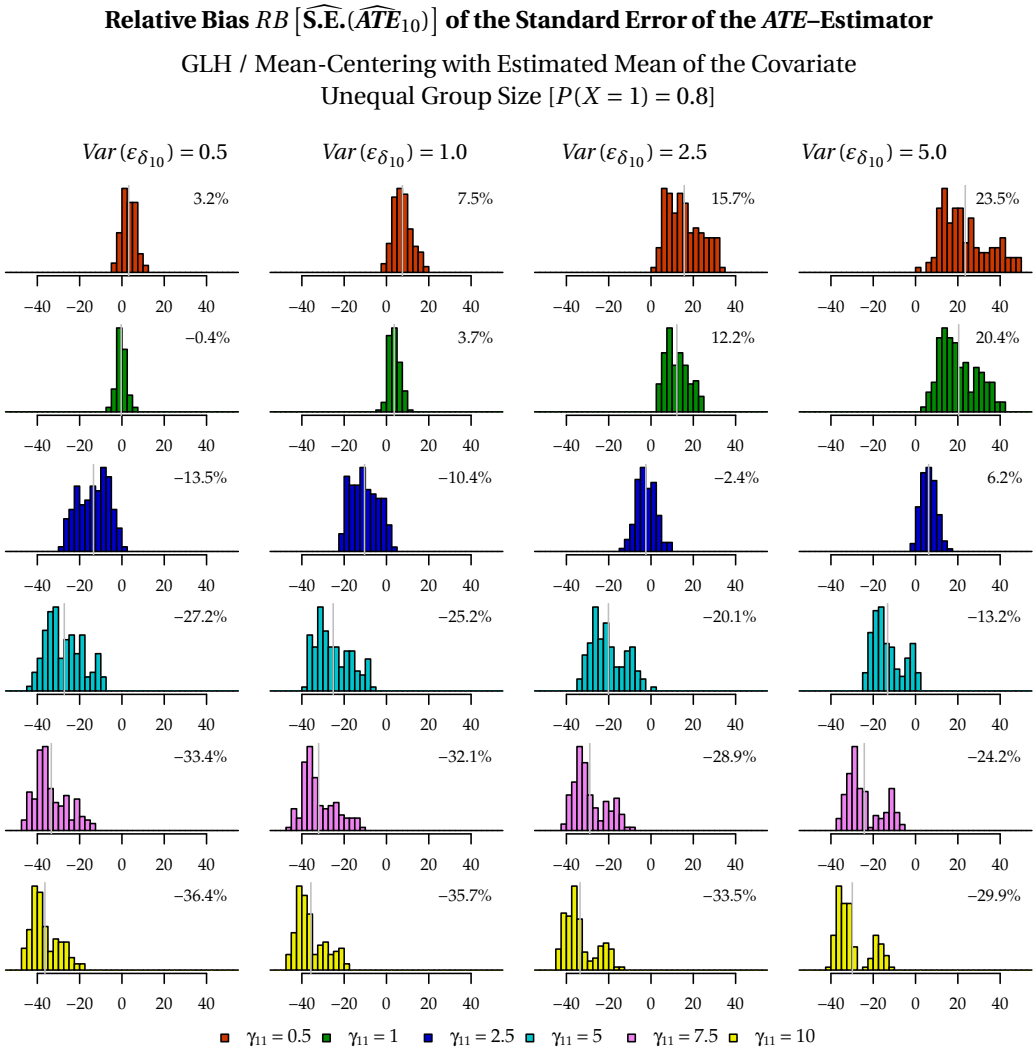


Figure 4.9: Relative bias of the standard error of the *ATE*–estimator: Histograms for the GLH / mean-centering based on the estimated mean of the covariate, conditional on γ_{11} and $Var(\varepsilon_{\delta_{10}})$ [$P(X = 1) = 0.8$]

orates the argument presented in section 3.2.5: The standard errors for the average total effect estimator obtained by ordinary least-squares are valid conditional on X and Z only. Increasingly underestimated standard errors are observable for increasing interaction effects (predicted as quadratic dependency), but this observation is valid conditional only on the level of dependency between X and Z [$R^2_{X|Z}$], on the group size [$P(X = 1)$], and on the additional factor heterogeneity of residual variance [manipulated as $Var(\varepsilon_{X=0})$ and as $Var(\varepsilon_{\delta_{10}})$ used for the data generation].[18] This line of argumentation will be clear when one follows

[18] A comparison of Figure 4.11 to the additional Figure 20 on page 31 of the digital appendix validates again that although the standard error of the average total effect estimator is unbiased for equal group sizes when the GLH / mean-centering approach is applied based on the true population value of the covariate, this result does not generalize with respect to the robustness of the GLH / mean-centering procedures when residual variances are heterogeneous and group sizes are unequal. Note that we also found an interesting dependency of the mean squared error of the standard error and the residual variances used for generating the data. This is presented

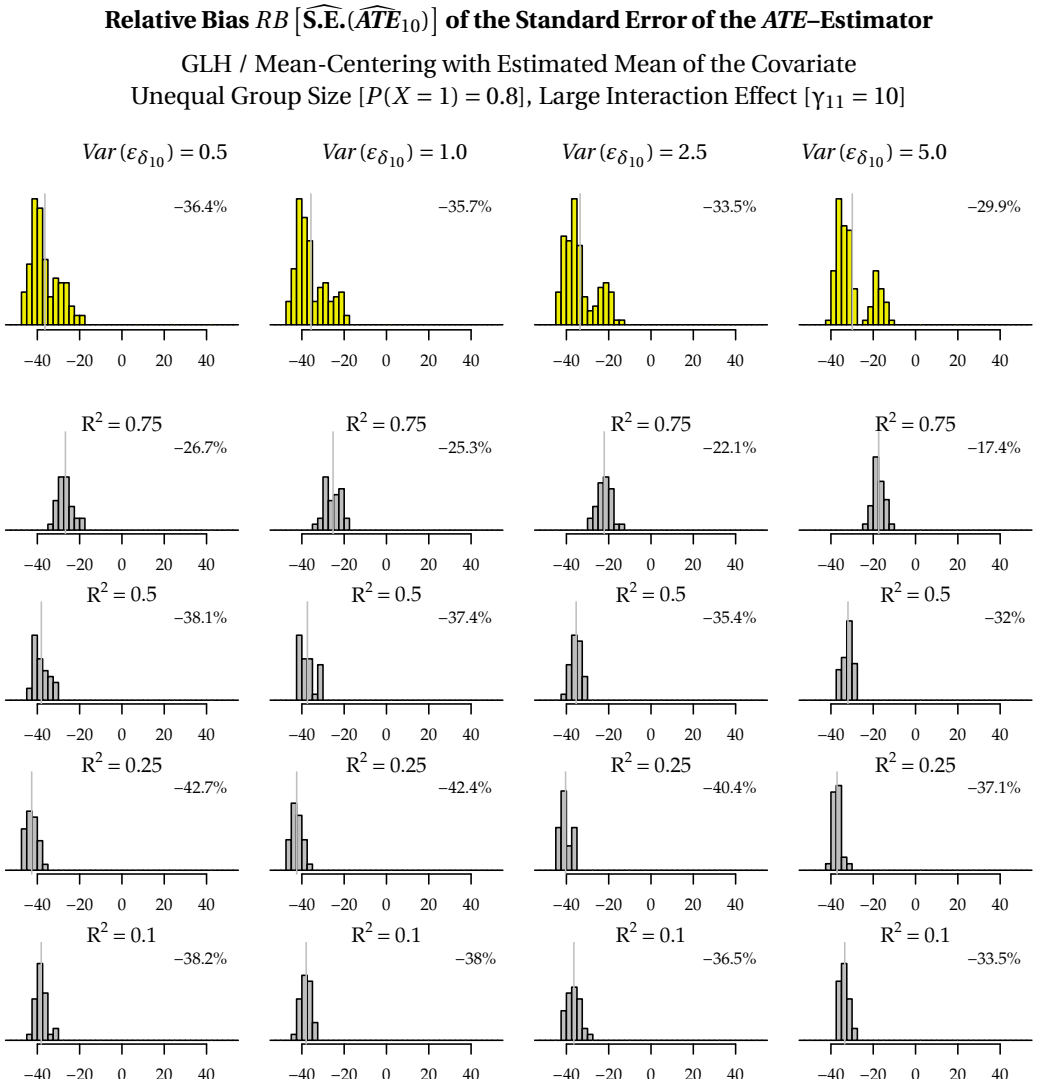


Figure 4.10: Relative bias of the standard error of the *ATE*–estimator: Histograms for the GLH / mean-centering based on the estimated mean of the covariate, conditional on $Var(\varepsilon_{\delta_{10}})$ and $R^2_{X|Z}$ [$P(X = 1) = 0.8$ and $\gamma_{11} = 10$]

the increase of $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ in Figure 4.11 for a fixed group size and for a fixed level of residual variances over the different conditions of the interaction parameter γ_{11}.

Summary The average total effect estimator is unbiased for all procedures based on the ordinary least-squares estimated covariate-treatment regression. The simulation study confirmed that test statistics for the hypothesis of no average total effect based on the general linear hypothesis and based on the moderated regression with mean-centered covariates produces (exactly) the same results. This conclusion is valid for

for a selected condition of the simulation study in the additional Figure 21 on page 32 of the digital appendix. Results for all simulated conditions are included in the digital supplement.

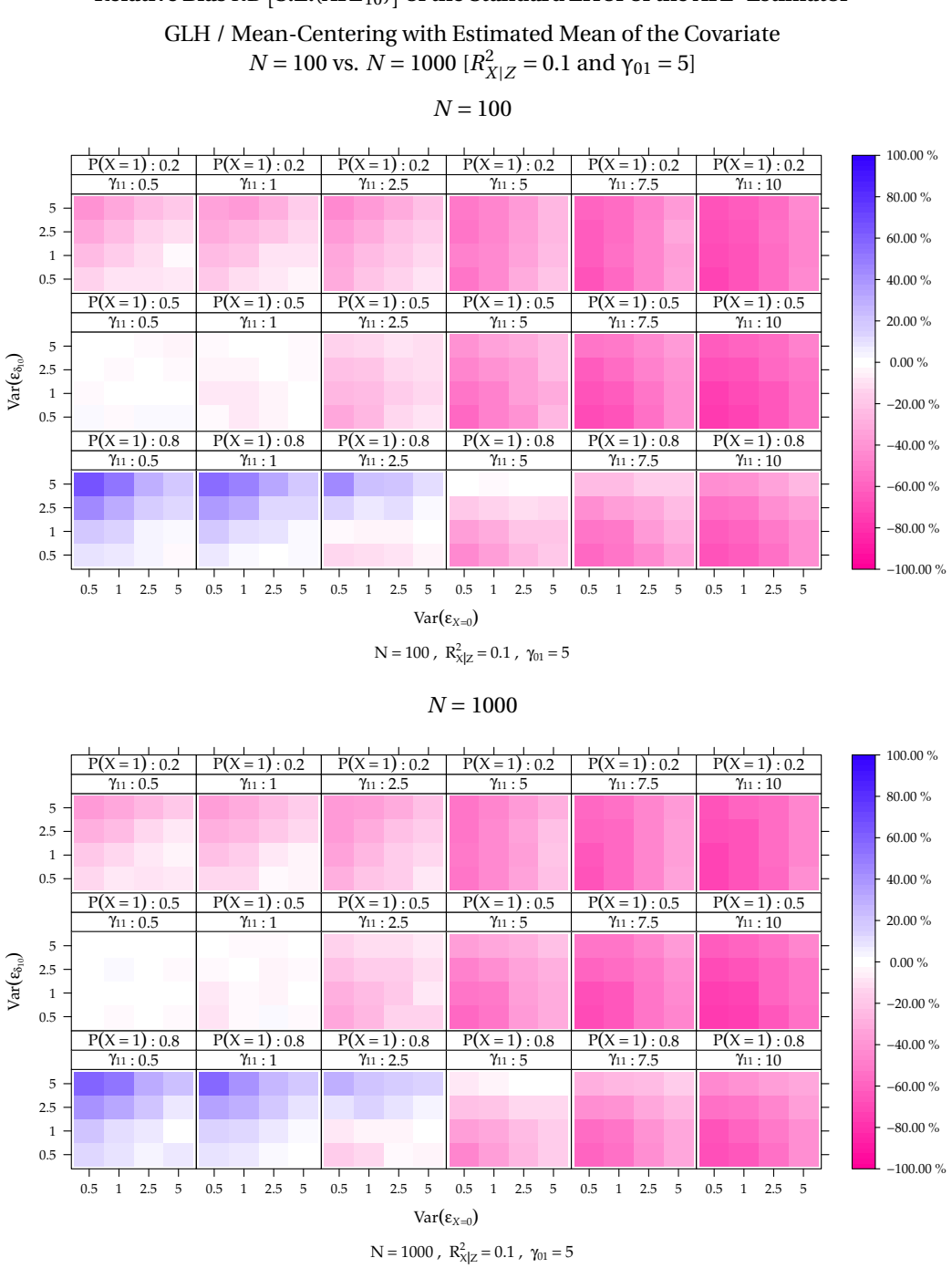


Figure 4.11: Relative bias of the standard error of the *ATE*–estimator: Level plots for the GLH / mean-centering based on the estimated mean of the covariate [$N = 100$ vs. $N = 1000$; $R^2_{X|Z} = 0.1$ and $\gamma_{01} = 5$]

the GLH / mean-centering approach with the covariate's estimated mean as well as for the test statistics using the population value of the covariate's expectation.

The $MSE\left[\widehat{ATE}_{10}\right]$ for the *ATE*–estimator for procedures based on the true expectation of the covariate increases as a function of the covariate-treatment interaction.

For equal group sizes, the empirical type-I-error rates are within the confidence bands around the nominal level for the test statistics incorporating the covariate's true expectation and for covariate-treatment regressions without interaction term. If regression lines are parallel, the GLH / mean-centering approach can be applied for equal group sizes. The general linear hypothesis (GLH) and the mean-centering approach generally fail for unequal group sizes, even for the studied conditions with only small violations of the assumption of homogeneity of residual variance and even if the covariate's true expectation is incorporated in the general linear hypothesis. That means that both procedures give clearly misleading results for testing hypotheses about the average total effect for unequal group sizes. [19]

We expected the unadjusted ordinary least-squares regression to give underestimated standard errors for the average total effect estimator, because the unconditional variance of the estimated average total effect is supposed to be underestimated for non-parallel regression lines in the covariate-treatment regression. The observed biased standard errors for the *ATE*–estimator impressively demonstrate the theoretical considerations presented in section 3.2.5.

In section 3.1.3 we summarized the literature dealing with stochastic regressors in moderated regression models. We pointed out that for covariate-treatment regressions with interaction terms the joint distribution of X, Z and Y needs to be considered. Accordingly, we discovered a relationship between the (relative) bias of the estimated standard error of the average total effect estimator and the dependency between the treatment variable and the covariate. The amount of confounding (manipulated as γ_{01}) was not directly connected to the underestimation of the standard error of the average total effect for conditions with non-parallel regression lines.

4.5.2 Heteroscedasticity Consistent Estimator

Heretofore, the results of the simulation study have reconfirmed that the empirical type-I-error rates of the GLH / mean-centering procedure are inflated for two different reasons: Firstly, because the standard errors obtained from the (conditional) ordinary least-squares estimators are not unconditionally valid if covariate-treatment interactions are present and covariates are generated as stochastic regressors. Secondly, because the standard errors for the different ordinary least-squares estimators are not robust against heteroskedasticity if group sizes are unequal. We now present the results for the same generated datasets obtained by applying one of the two heteroskedasticity consistent estimators (i. e., robust standard errors) described in

[19] The distributions of the observed rejection frequencies for all methods described in this section are presented as the additional Figure 41 on page 52 of the digital appendix.

subsection 3.2.2 (see page 58). It was expected that using robust standard errors should correct for violation of the homoskedasticity–assumptions of the general linear model. [20]

Type-I-Error Rate The observed empirical type-I-error rates for the adjusted GLH / mean-centering approach based on the estimated mean of the covariate are presented as level plots in Figure 4.12. The results for the same conditions of the simulation study for the uncorrected GLH / mean-centering approach have already been shown in Figure 4.7 and discussed in the previous subsection. A comparison of both level plots reveals that the `HC3` corrected test statistic yields acceptable type-I-error rates for small to medium interaction effects for unequal group sizes, i. e., under conditions where the unadjusted GLH / mean-centering approach had clearly inflated empirical type-I-error rates. The performance of the corrected test statistic is superior for conditions with strong dependency between X and Z (upper part of Figure 4.12).[21]

Bias of the Standard Error for the *ATE*–Estimator To explain the pattern of the empirical type-I-error rates obtained with the heteroskedasticity corrected variance-covariance matrix of parameter estimates, Figure 4.13 displays the corresponding level plots of the relative standard error bias of the *ATE*–estimator.

A comparison of the `HC3` corrected standard errors with the uncorrected standard errors from the ordinary least-squares estimated covariate-treatment regression (i. e., a comparison of Figure 4.11 for the GLH / mean-centering approach without correction and Figure 4.14 for the GLH / mean-centering approach with the `HC3` correction) demonstrates that the observed $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for unequal group sizes is caused by two different violations of the general linear models assumptions. The first two columns (for $\gamma_{11} = 0.5$ and $\gamma_{11} = 1$) in the upper part (for $N = 100$) and in the lower part (for $N = 1000$) of Figure 4.14 repeat the finding already described for the empirical type-I-error rates: Using a heteroskedasticity consistent estimator of the variance-covariance matrix of parameter estimates extends the robustness of the GLH / mean-centering approach against small interaction effects to conditions with unequal group sizes [$P(X = 1) = 0.2$ and $P(X = 1) = 0.8$]. Note that the remaining relative biases of the *ATE*–estimators' standard error for interaction effects of $\gamma_{11} > 1$ (i. e., the consequences of the violated fixed–X assumption) are not affected by the `HC3` correction and are obviously not connected to the sample size (compare the upper and the lower part of Figure 4.14).

A comparison of the two selected corrections (`HC3` versus `HC4`) yielded no significant results for large sample sizes.[22] For a small sample size ($N = 100$) we observed a noticeable over-correction of the `HC4`

[20] The absolute bias of the average total effect estimator and the mean squared error for the estimation of the average total effect are not reported again as these quantities are identical to the unadjusted GLH / mean-centering procedures.

[21] Note that there is again no observable difference between correcting the general linear hypothesis for heterogeneity of residual variances and the application of the corrected standard error for the centering approach in a moderated regression formulation of generalized analysis of covariance.

[22] Both approaches are expected to differ with respect to their small sample performance, see the additional Figure 22 on page 33 of the digital appendix.

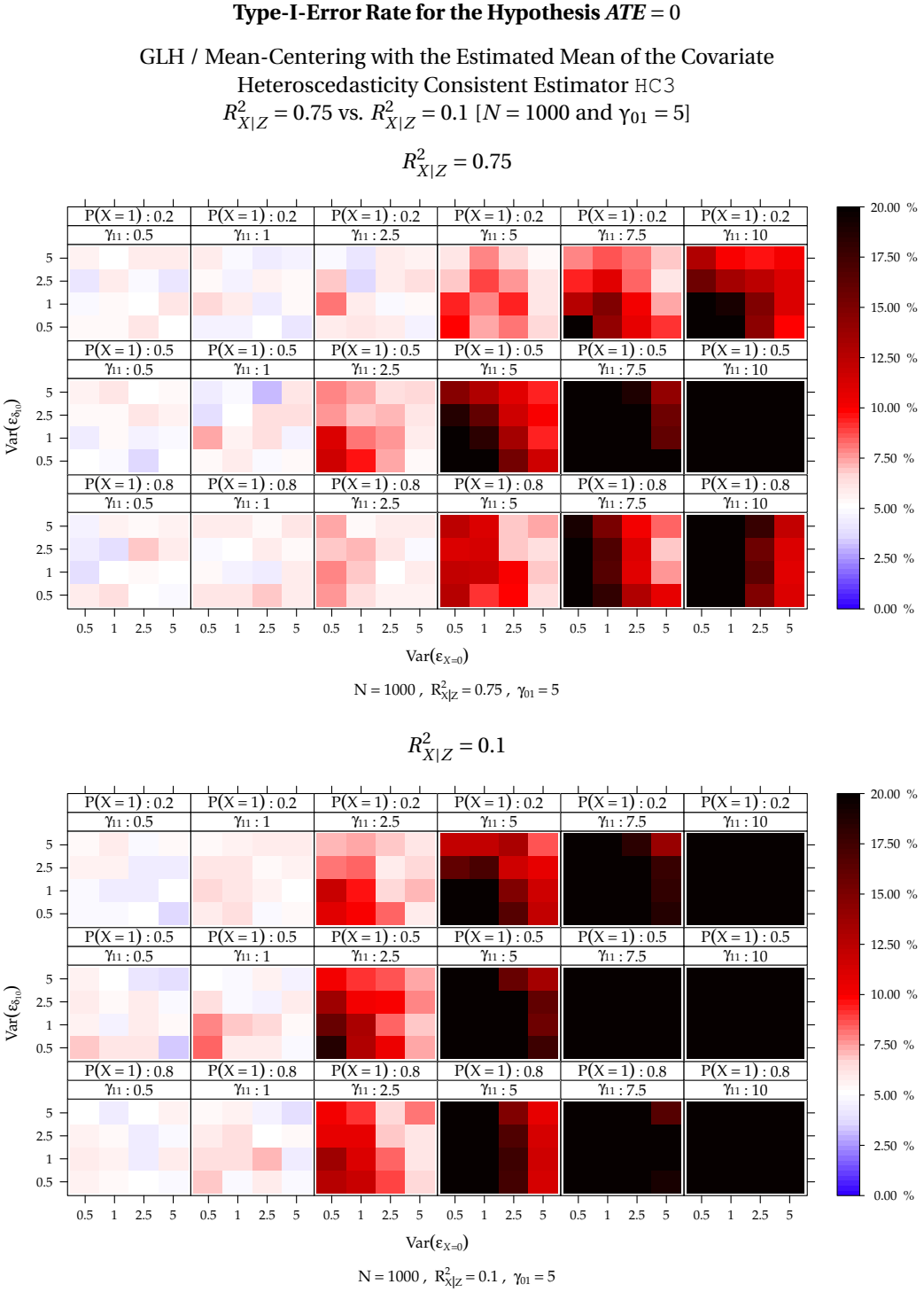


Figure 4.12: Type-I-error rate: Level plots for the GLH / mean-centering approach based on the estimated mean of the covariate, `HC3` corrected [$R^2_{X|Z} = 0.75$ vs. $R^2_{X|Z} = 0.1$; $N = 1000$ and $\gamma_{01} = 5$]

adjustment of the standard error for unequal group sizes. Due to this over-correction, the standard error

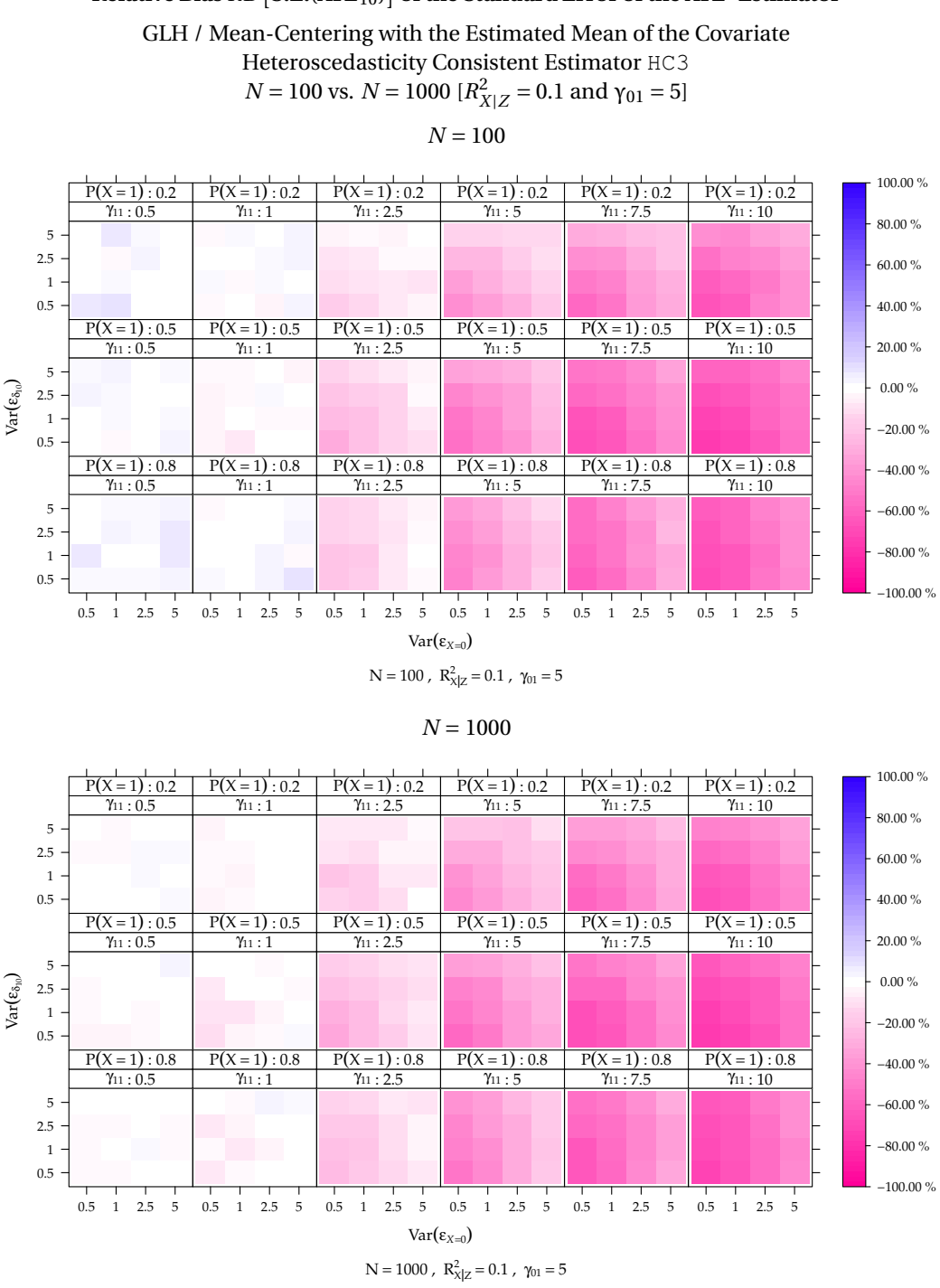


Figure 4.13: Relative bias of the standard error of the *ATE*–estimator: Level plots for the GLH / mean-centering approach based on the estimated mean of the covariate, `HC3` corrected [$N = 100$ vs. $N = 1000$; $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

of the *ATE*–estimator is overestimated for unequal group sizes and small sample sizes (see the first and the third row in the lower part of Figure 4.14).[23]

[23] We obtained unbiased estimates of the standard error for the average total effect estimator as well as correct empirical type-I-error rates for large sample sizes for the GLH / mean-centering approach based on the population value of the expectation of the covariate with `HC3` correction (see the additional Figure 23 on page 34, Figure 24 on page 35 and Figure 25 on page 36 of the digital appendix).

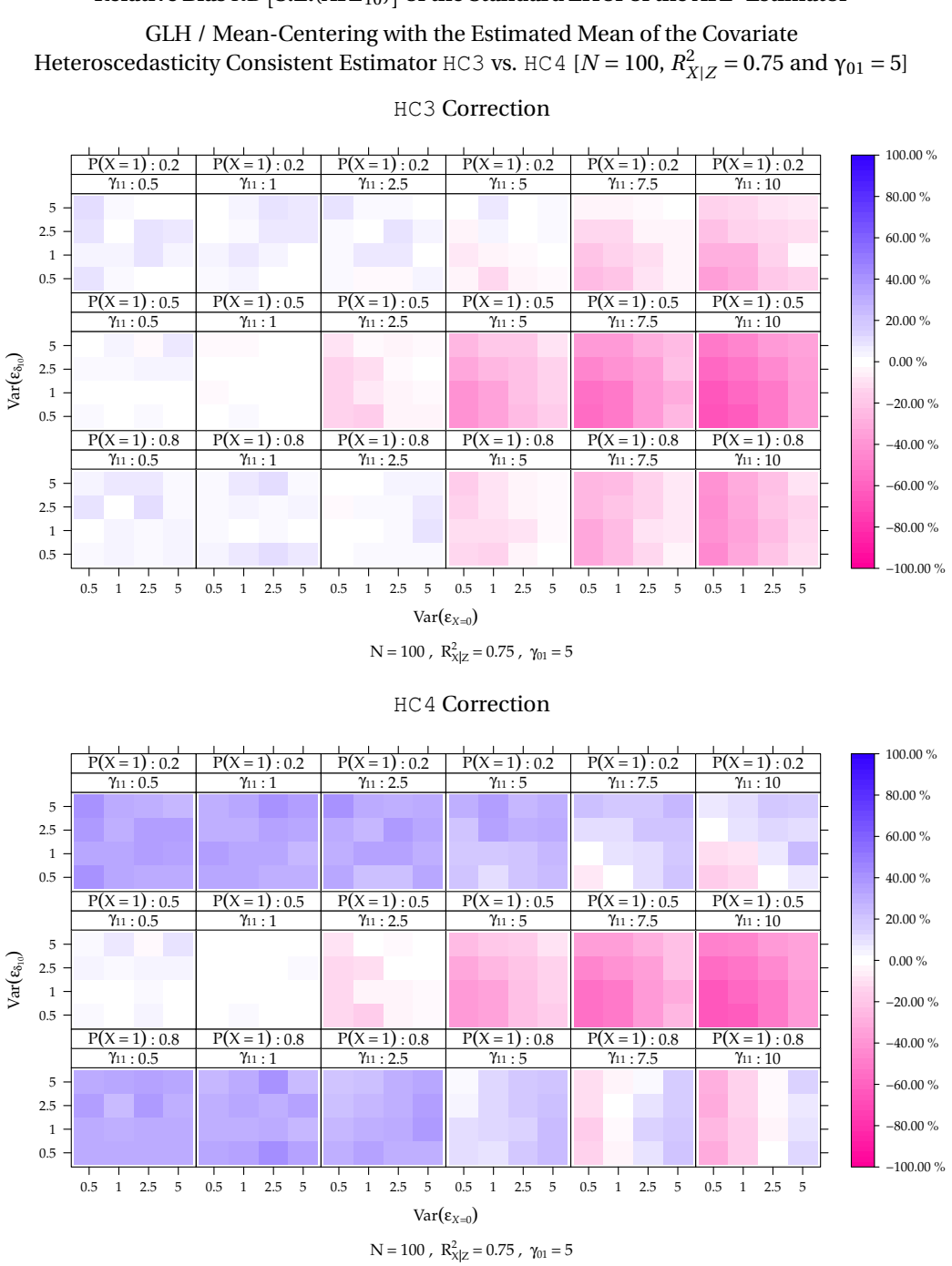


Figure 4.14: Relative bias of the standard error of the *ATE*–estimator: Level plots for the GLH / mean-centering approach based on the estimated mean of the covariate, HC3 corrected vs. HC4 corrected [$N = 100$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

Summary Robust standard errors based on a heteroskedasticity consistent estimators of the variance-covariance matrix reflect adequately the *ATE*–estimator's variability for covariate-treatment regressions without covariate-treatment interactions. The results for the two studied corrections HC3 and HC4 indicate that

heterogeneity of residual variance and the stochasticity of the covariate are distinct challenges when testing the hypothesis of no average total effect for data obtained from quasi-experimental designs. For equal group sizes, independent of whether a heteroskedasticity consistent estimator was applied or not, standard errors were underestimated for medium and strong covariate-treatment interactions.

Finally, a direct comparison of the two selected alternative heteroskedasticity consistent estimators favored the `HC3` correction because, under the conditions studied in the first part of the Monte Carlo simulation, the `HC4` method over-adjusted the standard errors of the *ATE*–estimators for conditions with small sample sizes and unequal group sizes.

4.5.3 Regression Estimates

In the following two subsections we describe the performance of the adjusted standard errors for the average total effect estimator based on *regression estimates* and based on *predictive simulations*. As for all other regression models with correctly specified mean-models (i. e., covariate-treatment regressions without first order misspecifactions, see Long & Trivedi, 1992), the *ATE*–estimator based on regression estimates and predictive simulations is unbiased.

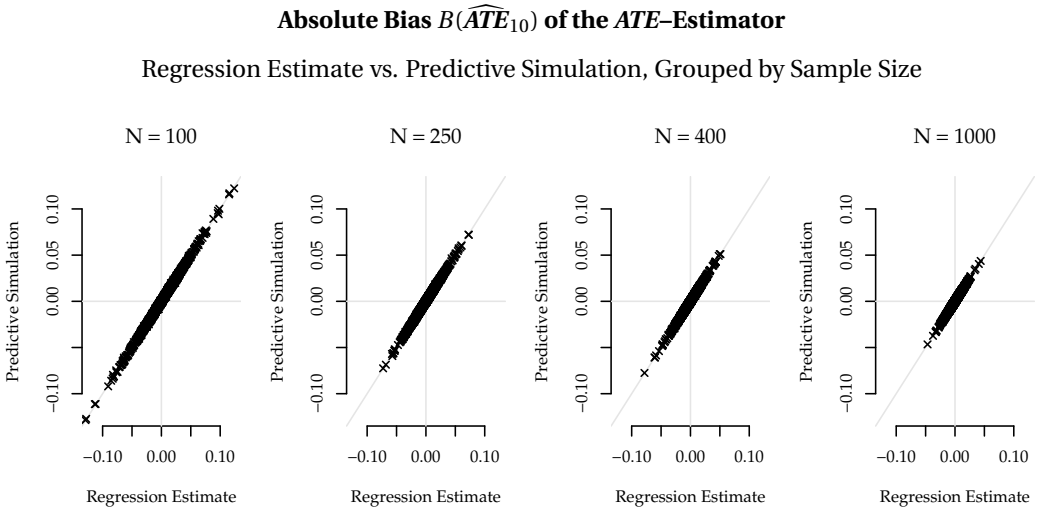


Figure 4.15: Absolute bias of the *ATE*–estimator: Scatter plots for a comparison of the regression estimates vs. predictive simulations, grouped by sample size N

Absolute Bias and Mean Squared Error of the *ATE*–Estimator A comparison of the *ATE*–estimator's absolute biases for regression estimates (x–axis) and predictive simulations (y–axis, presented in Figure 4.15) with the absolute biases of the GLH / mean-centering approach and the absolute biases of the multi-group structural equation models (presented in Figure 4.1 on page 117) reveal some minor random derivations of the average total effect estimator based on the predictive simulations which are due to the simulation

procedure. This additional small variability of the *ATE*–estimator does not depend on the amount of interaction.[24] We find no noteworthy differences between the regression estimates and the predictive simulation approach with respect to the mean squared error of the average total effect estimator (see Figure 4.16). The magnitude of the mean squared error decreases — as expected — with increasing sample sizes (notice the different scales within Figure 4.16) but this is true for both methods (all points are on the identity line).

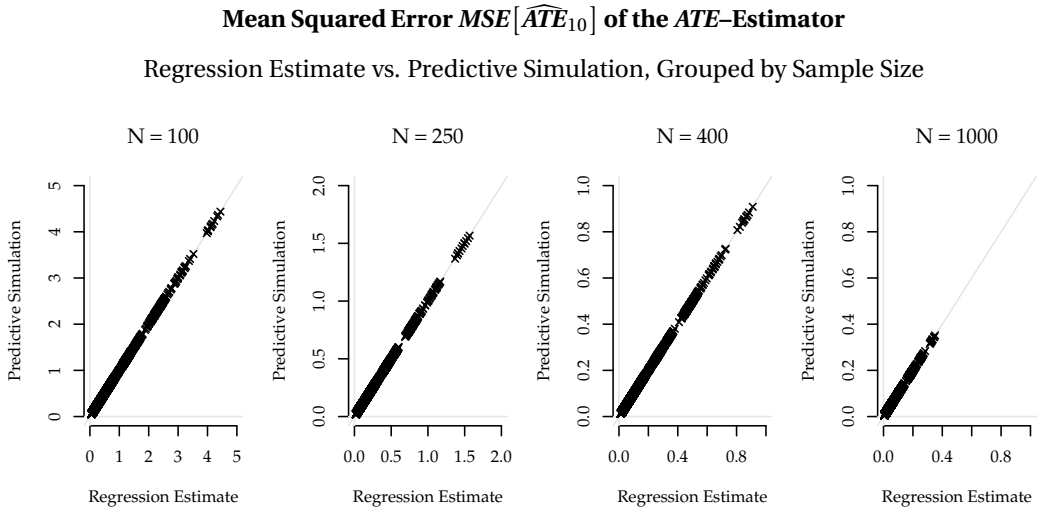


Figure 4.16: Mean squared error of the *ATE*–estimator: Scatter plots for a comparison of the regression estimates vs. predictive simulations, grouped by sample size N

Type-I-Error Rate The obtained empirical type-I-error rates for tests of the hypothesis $ATE = 0$ based on the regression estimates with adjusted standard errors as suggested by Schafer and Kang (2008) are surprisingly good for equal group sizes: Figure 4.17 compares the normal approximation (z–test, x–axis) and the t–test (y–axis) for all conditions of the simulation study I with $P(X = 1) = 0.5$ (presented as separate scatter plots for each level of the interaction parameter γ_{11} used for data generation). The rejection frequencies obtained from both test statistics for almost all studied conditions with equal group sizes are within the confidence bands (marked by gray and red horizontal and vertical lines), regardless of the interaction parameter γ_{11}.[25]

Two phenomena can be observed for conditions with unequal group sizes: On the one hand, inflated empirical type-I-error rates for small sample sizes ($N = 100$) are obvious.[26] On the other hand, it is interesting to notice the (slightly) different behavior of the t–test compared to the normal approximation based

[24] A direct comparison of the *ATE*–estimator obtained from the predictive simulations and the *ATE*–estimator obtained from the regression estimates is included in the digital appendix as additional Figure 26 on page 37.

[25] Furthermore, the empirical distribution of rejection frequencies, which is provided as the additional Figure 27 on page 38 of the digital appendix, reveals that there are no observable systematic small sample differences between the z–test and the t–test for conditions with equal group sizes.

[26] Note that this inflation is more obvious for datasets generated with small interaction effects for conditions with unequal group sizes (see the additional Figure 28 on page 39 and Figure 29 on page 40 of the digital appendix).

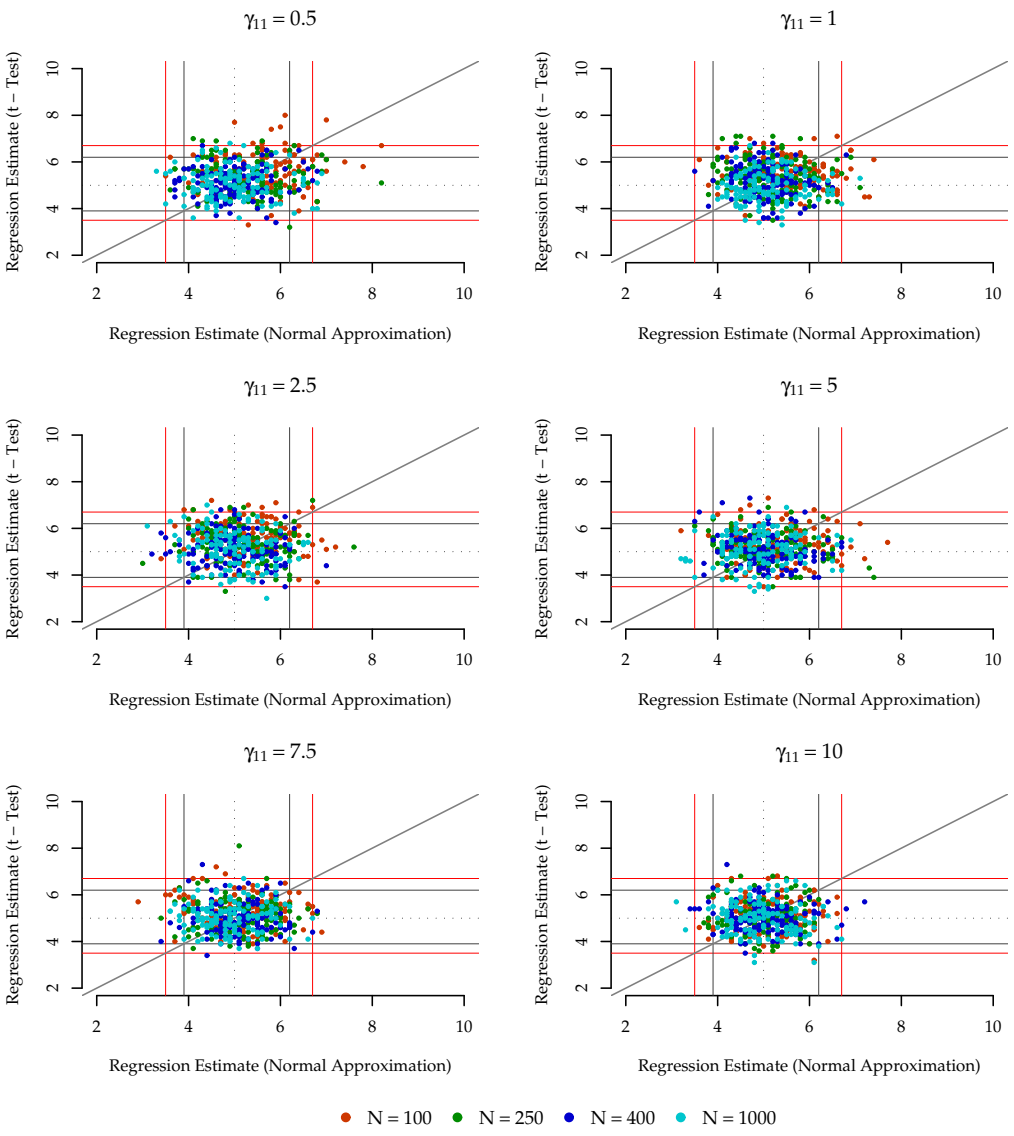


Figure 4.17: Type-I-error rate: Scatter plots for a comparison of the regression estimates based on a normal approximation and based on a t–test, grouped by interaction γ_{11} $[P(X = 1) = 0.5]$

z–test (in particular for conditions with strong interaction effects and small sample sizes). Whereas the distribution of the rejection frequencies for the z–test is closer to the desired symmetric distribution around the nominal 5 % level than is the distribution of rejection frequencies for the t–test for conditions of the simulation study I with $N = 100$ and $P(X = 1) = 0.2$, the reverse is true for conditions of the simulation

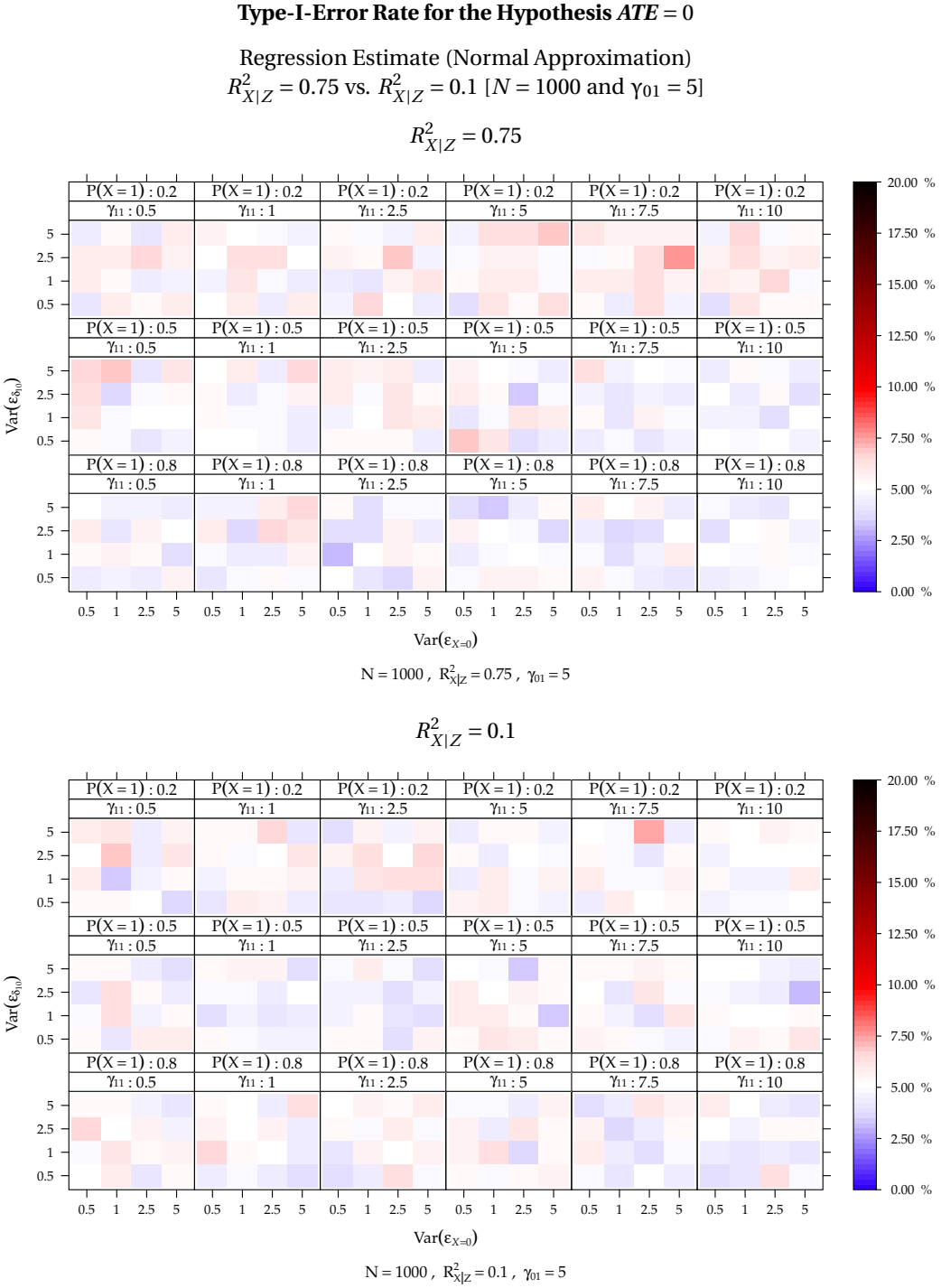


Figure 4.18: Type-I-error rate: Level plots for the regression estimates based on the normal approximation [$R^2_{X|Z} = 0.75$ vs. $R^2_{X|Z} = 0.1$; $N = 1000$ and $\gamma_{01} = 5$]

study with $P(X = 1) = 0.8$:[27] Here the normal approximation is worse for small sample sizes ($N = 100$). The rejection frequencies of the z–test and the t–test are identical for large sample sizes ($N = 1000$).[28]

[27] This is obvious from a comparison of the upper part of the additional Figure 30 on page 41 of the digital appendix to the lower part of the same figure.

[28] See also the additional Figure 31 on page 42 of the digital appendix for the corresponding level plots.

Figure 4.18 summarizes the findings for the regression estimates (z–test) regarding the empirical type-I-error rates for $N = 1000$, $\gamma_{01} = 5$ and two different values of the dependancy between X and Z as level plots for equal and unequal group sizes. A comparison of this figure with the corresponding figures for the GLH / mean-centering approach (see Figure 4.6 and Figure 4.7), as well as a comparison to the level plots generated from the results of the heteroscedasticity-consistent estimators (see, for instance, Figure 4.12) reveal that there is no observable systematic inflation of the empirical type-I-error rates for the regression estimates due to heterogeneity of residual variance (and unequal group sizes), or due to the amount of the covariate-treatment interaction.

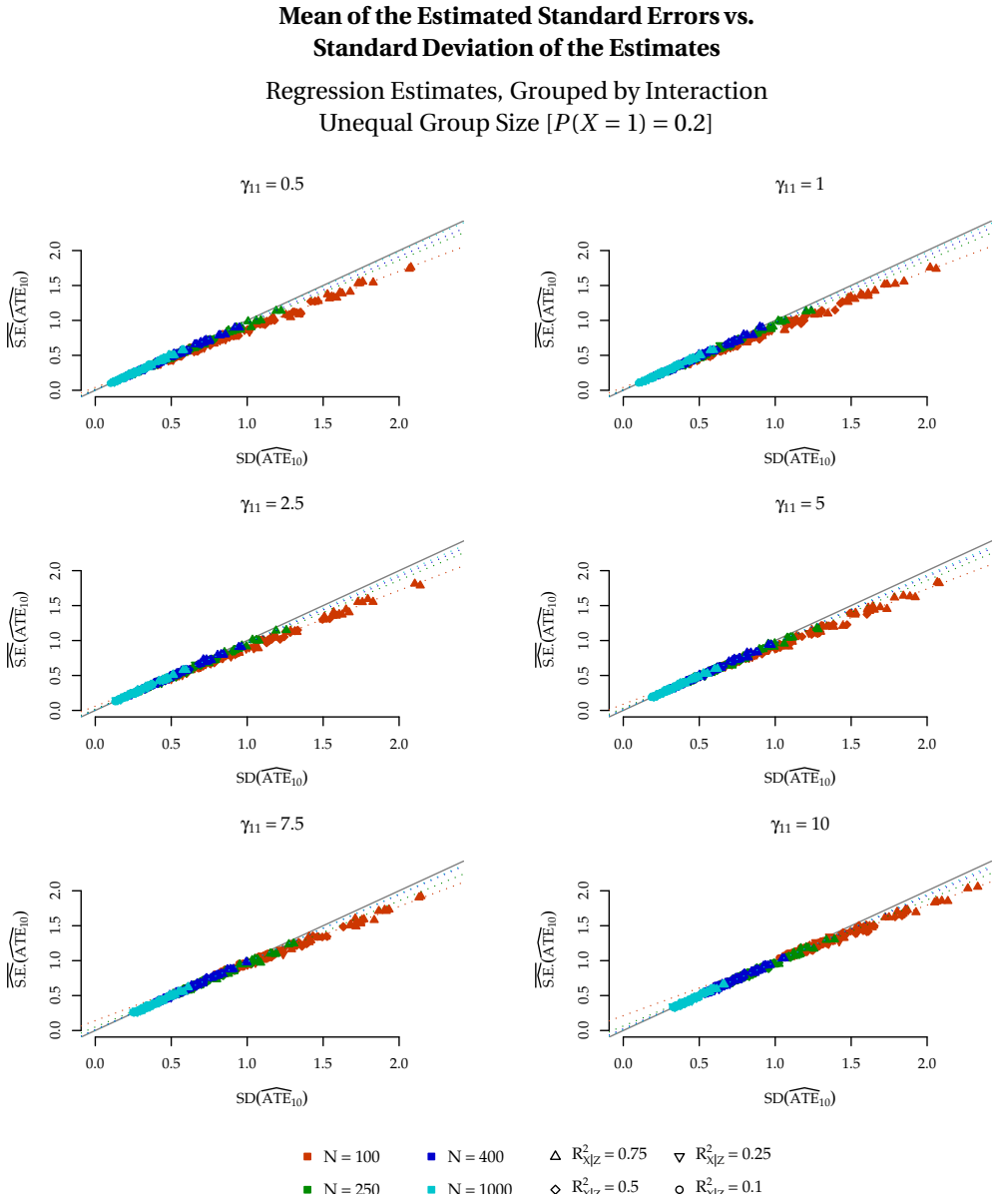


Figure 4.19: Mean of the estimated standard errors vs. standard deviation of the estimated average total effects, regression estimates, grouped by interaction γ_{11} [$P(X = 1) = 0.2$]

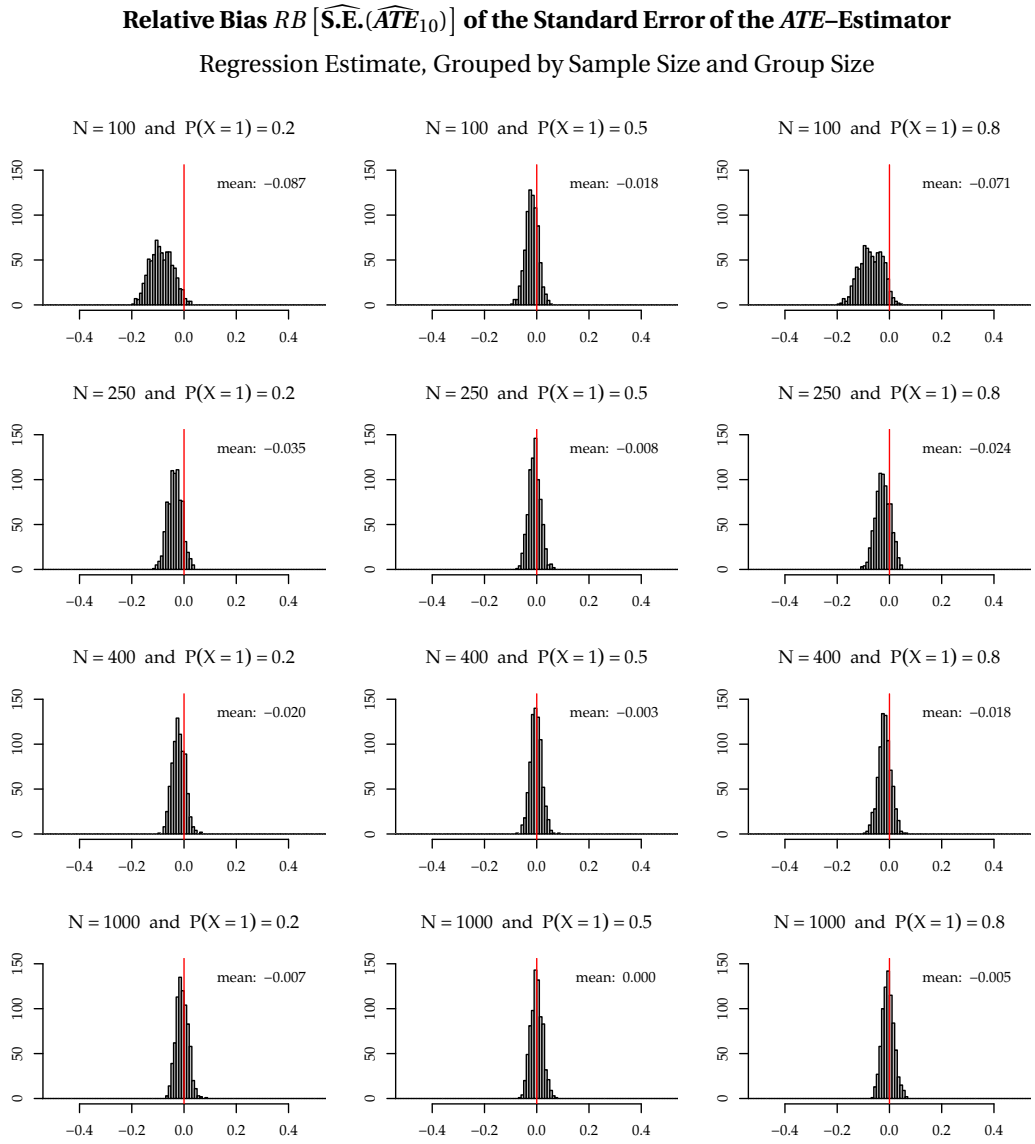


Figure 4.20: Relative bias of the standard error of the *ATE*–estimator: Histograms for the regression estimates, grouped by sample size N and group size $P(X = 1)$

Bias of the Standard Error of the *ATE*–Estimator The standard errors for the average total effect estimator corrected with the formulas given by Schafer and Kang (2008) are unbiased for almost all conditions of simulation study I. This is displayed in Figure 4.19 for unequal group sizes [$P(X = 1) = 0.2$]. Each symbol in the six scatter plots represents the empirical standard derivation of the *ATE*–estimates for one condition of the simulation study on the x–axis [i. e., $SD\left(\widehat{ATE}_{10}\right)$] and the average of the calculated standard errors for the *ATE*–estimator on the y–axis [i. e., $\overline{\widehat{\text{S.E.}}}\left(\widehat{ATE}_{10}\right)$] for the same condition of the simulation study's design. The dotted lines (in colors corresponding to the different sample sizes used for generating the datasets) show the results of a simple linear regression of the averaged standard errors on the standard deviation. For

equal group sizes the plotted regression lines differ only slightly from the diagonal line, meaning that the observer variability of the standard error is almost unbiasedly estimated with the adjusted standard errors.[29] For unequal sample sizes where the treatment group is smaller than the control group [$P(X = 1) = 0.2$], the relative bias $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ is apparently larger.[30]

The conditional distributions of the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for the regression estimate approach, approximated as histograms, are presented in Figure 4.20, grouped by sample size and group size. Obviously, the standard error is biased for unequal group sizes and small sample sizes, but this bias vanishes if the sample size increases.[31]

Summary Surprisingly good results were observed for the adjusted standard errors of the *ATE*–estimator obtained from the ordinary least-squares estimated regression estimates. Although Schafer and Kang (2008) did not explicitly mention stochasticity of covariates in their derivation of the formulas for the adjusted standard errors, we verified the appropriateness of these standard errors empirically for the conditions studied in part I of the Monte Carlo simulation.

4.5.4 Predictive Simulation

We have already reported that the average total effect estimator obtained from the predictive simulation approach is unbiased (see paragraph 3.2.3 on page 61). The computed measures of efficiency for the estimation of the average total effect, i. e., the mean squared errors, were comparable for the predictive simulation approach and the regression estimate approach as well.[32]

Type-I-Error Rate The rejection frequencies for tests of the hypothesis $ATE = 0$ based on predictive simulations as suggested by Gelman and Hill (2007) are clearly higher than the nominal 5 % level for all studied conditions with serious interaction effects (see the distribution of the rejection frequencies over all conditions of simulation study I for equal group sizes in Figure 4.21). Obviously, this is not an effect of small sample sizes, because the asymmetric distribution of observed rejection frequencies does not even disappear for conditions with large sample sizes (see the right tails in the distributions for $N = 400$ and $N = 1000$ in Figure 4.21). Although the average total effect estimator is unbiased, the test statistic (the t-test as well as the normal approximation) produce acceptable rejection frequencies only for small interaction effects $\gamma_{11} \leq 1$, and only for simulated conditions with equal group sizes.

[29] As Figure 4.19 also shows, the absolute bias is largest for conditions with the highest value of the dependency between X and Z ($R^2_{X|Z} = 0.75$) and the smallest sample sizes ($N = 100$).
[30] See the additional Figure 32 on page 43 and Figure 33 on page 44 of the digital appendix.
[31] See the additional Figure 34 on page 45 of the digital appendix.
[32] See Figure 4.15 and Figure 4.16 above as well as the additional Figure 26 on page 37 of the digital appendix.

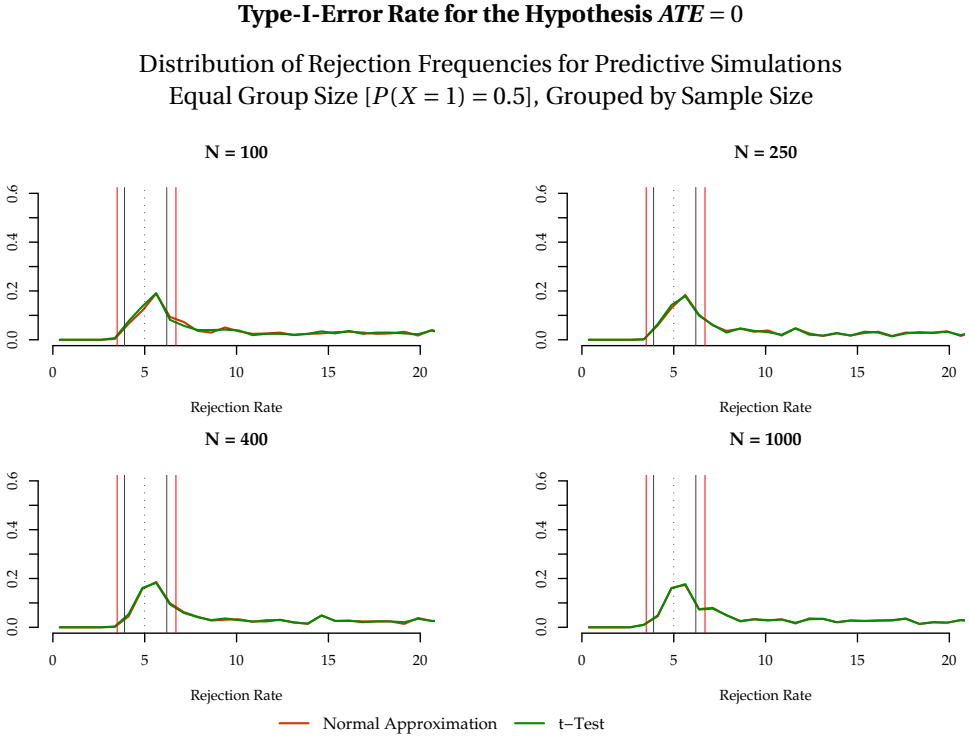


Figure 4.21: Type-I-error rate: Distribution of the rejection frequencies for the predictive simulation approach (normal approximation vs. t–test), grouped by sample size N [$P(X = 1) = 0.5$]

A direct comparison of the empirical type-I-error rates obtained for tests of the hypothesis of no average total effect based on the (unadjusted) GLH / mean-centering approach to the empirical type-I-error rates observed for the test statistic based on the standard errors obtained by predictive simulation (t–test)[33] is presented as level plot in Figure 4.22. Obviously, the simulation-based procedure produces only very small improvements of the empirical type-I-error rate. These benefits are likely to vanish when the power of the resulting test statistic (instead of the type-I-error rate) is considered.

Bias of the Standard Error of the *ATE*–Estimator The distributions of $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for the predictive simulation approach, approximated as histograms, are summarized in Figure 4.23, grouped by sample size and group size. For equal group sizes [$P(X = 1) = 0.5$] and for treatment groups smaller than control groups [$P(X = 1) = 0.2$], standard errors are strongly underestimated. For some conditions with a treatment probability of $P(X = 1) = 0.8$, the effect due to heterogeneity of between-group residual variances and the consequences of the inappropriately handled stochasticity of the covariates cancel each other out.[34]

[33] In contrast to the regression estimation approach discussed in the previous subsection, no meaningful difference between the t-test and the normal approximation can be observed. As the additional Figure 35 on page 46, Figure 36 on page 47 and Figure 37 on page 48 of the digital appendix demonstrate, the observed inflation of the empirical type-I-error rates are directly connected to the values of the interaction parameter γ_{11} used for generating the data.

[34] This is obvious from a comparison of the results presented in the additional Figure 38 on page 49, Figure 39 on page 50 and Figure 40 on page 51 of the digital appendix.

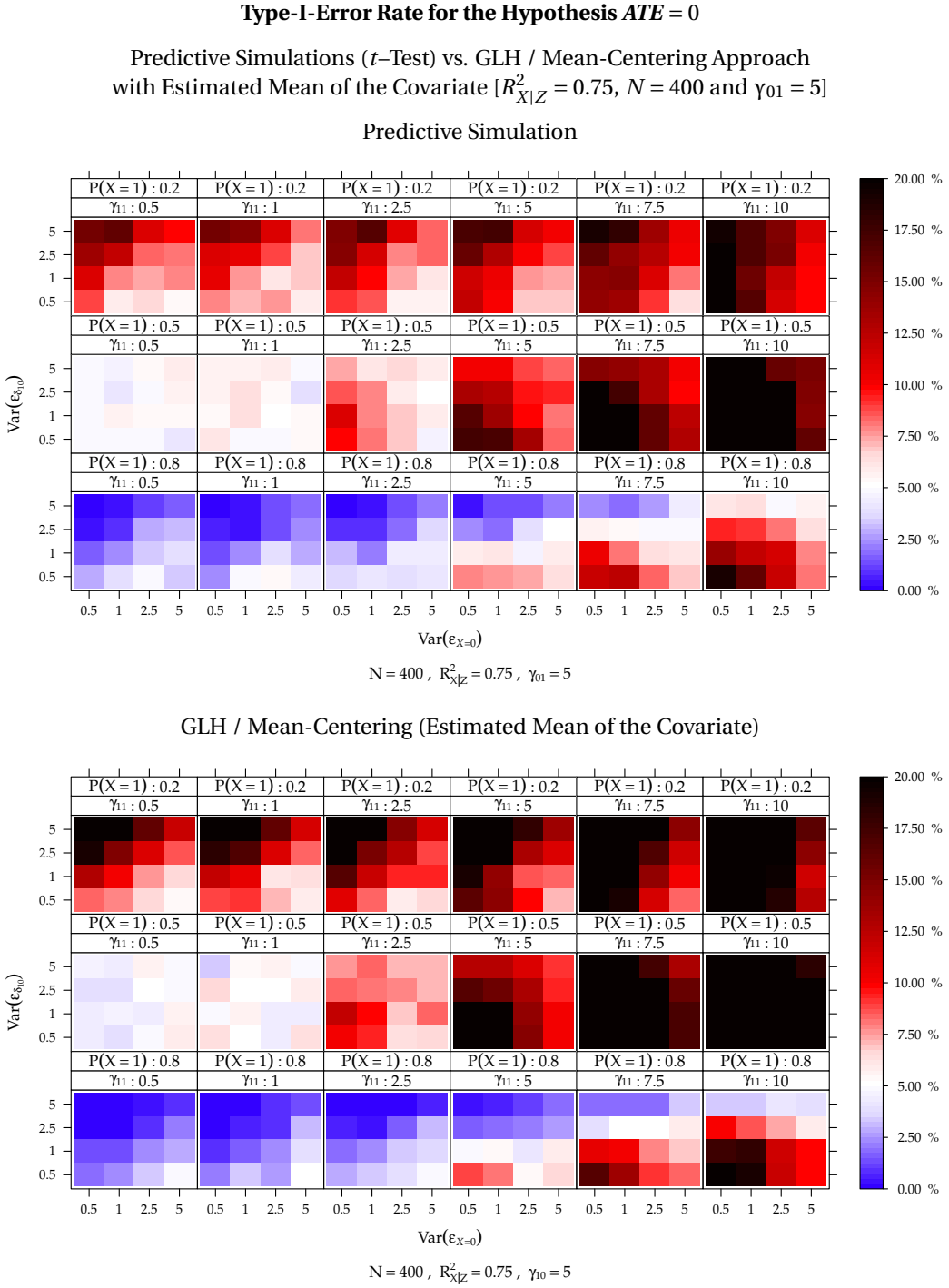


Figure 4.22: Type-I-error rate: Level plots for the predictive simulation approach vs. the GLH / mean-centering approach (estimated mean of the covariate) [$R^2_{X|Z} = 0.75$, $N = 400$ and $\gamma_{01} = 5$]

Summary The predictive simulation approach did not differ substantially from the GLH / mean-centering approach. Although Gelman and Hill (2007) suggest the simulation-based procedure for inference about nonlinear predictions, the unconditional variance of the *ATE*–estimator is underestimated for equal group

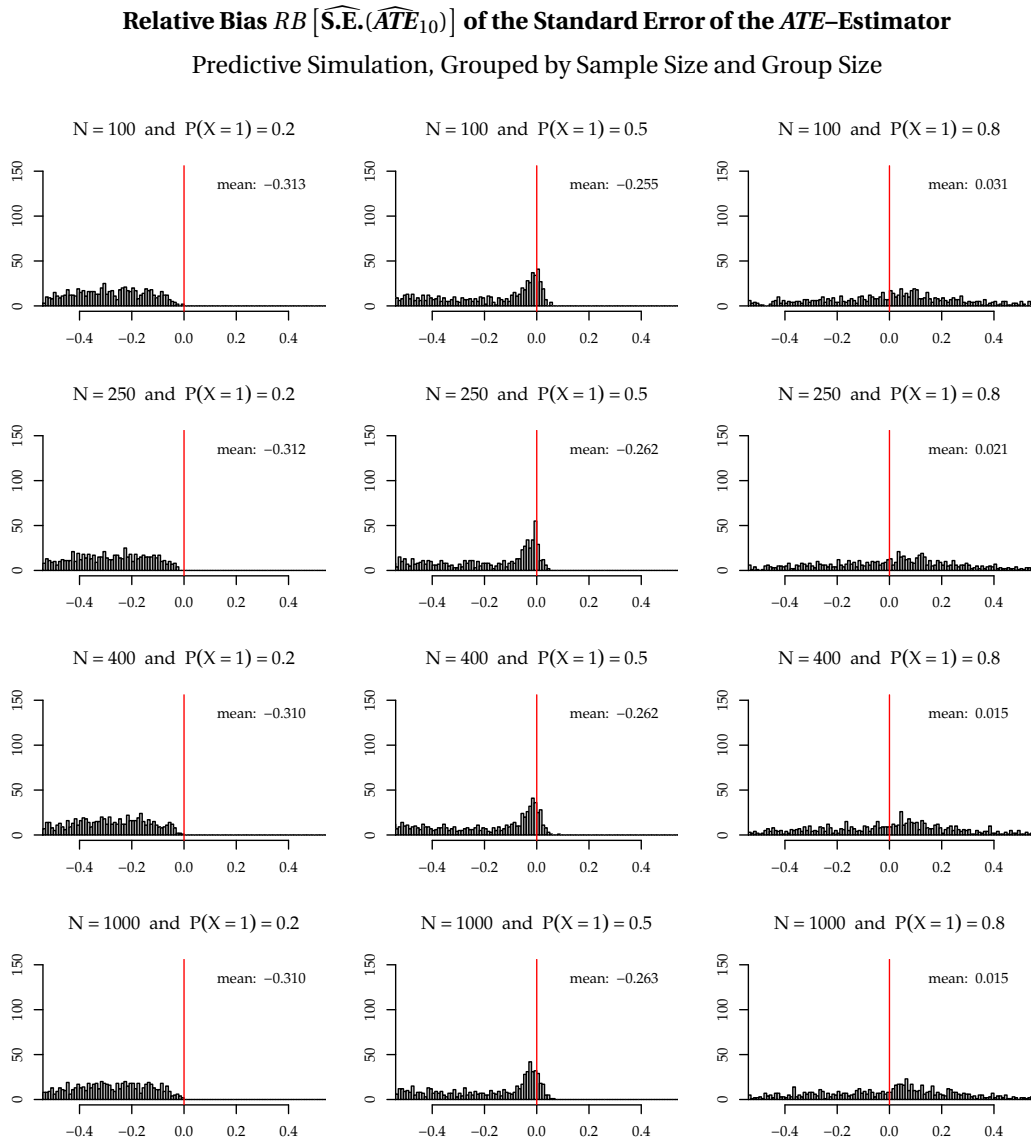


Figure 4.23: Relative bias of the standard error of the *ATE*–estimator: Histograms for the predictive simulation approach, grouped by sample size N and group size $P(X = 1)$

sizes and unequal group sizes with $P(X = 1) = 0.2$ and biased for $P(X = 1) = 0.8$. Hence, the predictive simulation approach will not be included in the second part of the simulation study.

4.5.5 Summary

In line with Flory (2004) and Nagengast (2006) we found inflated empirical type-I-error rates for tests of the hypothesis $ATE = 0$ obtained from the general linear hypothesis. Furthermore, we demonstrated that mean-centering of covariates with estimated means does not change the statistical properties of the *ATE*–estimator. In contrast to previous simulation studies, we manipulated the individual (total) effect's variability as an additional factor in the simulation design. We were thus able to disentangle the consequences

of covariates stochasticity (in conjunction with covariate-treatment interactions) and the consequences of heterogeneity of between-group residual variances (in conjunction with unequal group sizes). The general linear model was found to be neither robust against heteroscedasticity nor gainst violations of the fixed-X assumption. Nevertheless, the heteroscedasticity-consistent estimators were found to improve the standard errors of the *ATE*–estimators. The GLH / mean-centering approach based on these adjusted standard errors can be suggested for unequal group sizes and parallel regression slopes. For non-parallel regression slopes, the adjusted standard errors for regression estimates are the only recommendable approach to test hypotheses about average total effects. The recently published adjusted standard errors for regression estimates, as an alternative to the analysis of covariance, will be used as the benchmark for the performance of the structural equation models with non-linear constraints presented in the next two sections.

4.6 Results for the Structural Equation Models under Homogeneity of Residual Variance

The results for the different implementations of generalized analysis of covarinace as structural equation model with nonlinear constraints will be presented in the following two sections. We will start with a subset of the conditions studied in simulation study I, where the between-group residual variances are almost homogenous. Flory (2008) studied only conditions with variance homogeneity. Hence, in order to replicate Flory's findings we resrict the results included in this first section to datasets generated under similar conditions.

4.6.1 Simple Multi-Group Model

The crucial point in the application of the *simple multi-group model* as previously described is the assumption of a known treatment probability (see page 83). Inflated empirical type-I-error rates and biased standard errors of the *ATE*–estimator are expected if the true treatment probability in the nonlinear constraint is replaced by the estimated value of the group size (i. e., if the estimated mean of the treatment variable is incorporated in the nonlinear constraint). Hence, we investigate the observed robustness of the simple multi-group model against the stochasticity of X in this subsection (see reseach question in subsection 3.4.4).

Type-I-Error Rate The rejection frequencies for the Wald–tests of the hypothesis $ATE = 0$ are presented in this paragraph. For this test statistic, the inverse of the asymptotic variance-covariance matrix of the nonlinear constraint's parameters is pre- and post-multiplied with the value of the constraint itself (see page 78). Therefore, the overall observed rejection frequencies can be understood as an aggregated description of the performance of this implementation of generalized analysis of covariance with respect to the abso-

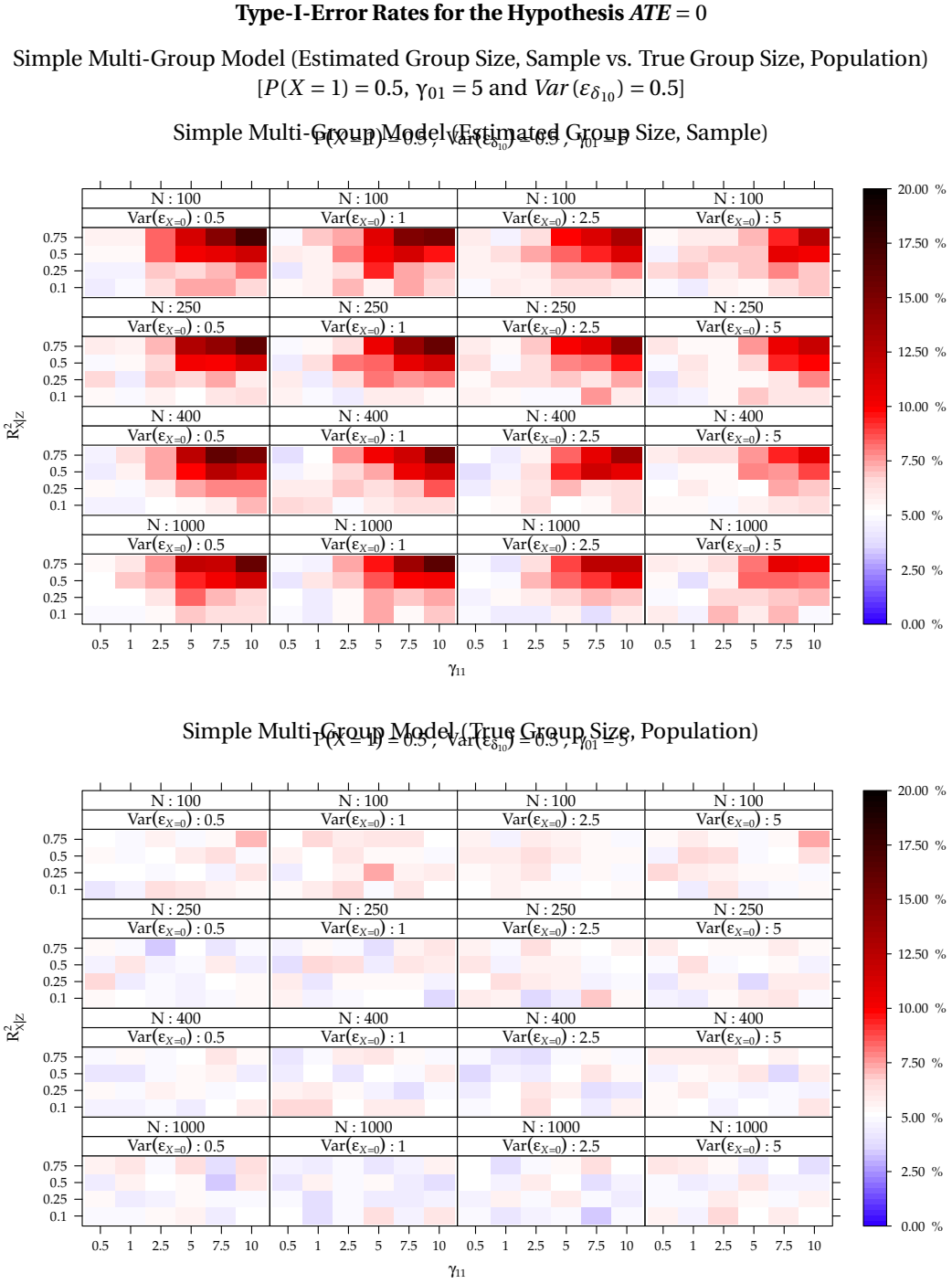


Figure 4.24: Type-I-error rate: Level plots for a comparison of the simple multi-group model based on estimated group size (sample) and the simple multi-group model based on the true group size (population) $[P(X = 1) = 0.5, \gamma_{01} = 5$ and $Var(\varepsilon_{\delta_{10}}) = 0.5]$

lute bias $B(\widehat{ATE}_{10})$ of the estimator (i. e., the value of the nonlinear constraint), and the relative bias of the standard error $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ of the *ATE*–estimator.

Empirical type-I-error rates for the simple multi-group model based on the estimated group size (sample), as well as for the simple multi-group model based on the true population value of the group size (population) are displayed in Figure 4.24 for equal group sizes.[35] The empirical type-I-error rates for the model based on the sample estimate of the group size (upper part of the figure) are clearly inflated for all simulated conditions with $\gamma_{11} > 1$ and a dependency between X and Z of $R^2_{X|Z} > 0.1$, even for almost homogenous between-group residual variances, i. e., $Var(\varepsilon_{X=0}) = 5$ (see the forth outer column of the level plot).[36] The empirical type-I-error rates are inflated due to the estimated group size, i. e., the stochasticity of X (compare the upper and the lower part of Figure 4.24).[37] The systematic pattern of inflated empirical type-I-error rates for medium and strong interaction effects and substantial dependencies of X and Z are not observed if the multi-group model is implemented with the true group size in the nonlinear constraint (see the lower part of Figure 4.24).

Table 4.6: Absolute bias of the *ATE*–estimator: Comparison of the simple multi-group models (population vs. sample) [$P(X = 1) = 0.5$, $\gamma_{01} = 5$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $Var(\varepsilon_{X=0}) = 0.5$]

		Interaction (γ_{11})					
$R^2_{X\|Z}$	N	0.5	1.0	2.5	5.0	7.5	10.0
Simple Multi-Group Model (Population)							
0.25	100	−0.0029	−0.0027	−0.0013	−0.0040	0.0064	**0.0109**
	250	0.0011	0.0008	0.0091	**-0.0144**	0.0023	**0.0129**
	400	0.0058	0.0005	−0.0004	−0.0024	−0.0037	−0.0041
	1000	0.0028	0.0021	−0.0010	0.0034	0.0006	0.0009
0.75	100	0.0006	**0.0131**	0.0048	**-0.0150**	0.0012	**-0.0215**
	250	0.0043	0.0014	−0.0001	**-0.0112**	**-0.0213**	0.0050
	400	−0.0027	0.0058	0.0029	0.0024	**-0.0183**	**-0.0318**
	1000	−0.0046	−0.0054	0.0012	−0.0092	**-0.0110**	**-0.0250**
Simple Multi-Group Model (Sample)							
0.25	100	−0.0028	−0.0069	0.0014	−0.0018	**0.0112**	0.0068
	250	0.0014	0.0005	0.0069	−0.0022	0.0021	0.0002
	400	0.0054	0.0002	0.0001	−0.0017	−0.0077	−0.0004
	1000	0.0028	0.0023	−0.0014	0.0018	−0.0011	−0.0017
0.75	100	0.0013	**0.0163**	**0.0126**	−0.0053	0.0076	−0.0090
	250	0.0045	0.0034	0.0084	−0.0090	0.0016	0.0066
	400	−0.0027	0.0066	0.0041	0.0070	0.0003	−0.0001
	1000	−0.0040	−0.0021	0.0060	0.0030	−0.0005	0.0022

[35] Note that compared to the level plots presented so far, the results in Figure 4.24 are structured in a different way: For a fixed treatment probability $P(X = 1) = 0.5$, a fixed level of the factor $Var(\varepsilon_{\delta_{10}}) = 0.5$, and for a fixed amount of a confounding $\gamma_{01} = 5$ the observed type-I-error rates are condensed into one single plot.

[36] This inflation of the empirical type-I-error rate is slightly more conspicuous for conditions with smaller residual variances of the outcome model, i. e., $Var(\varepsilon_{X=0}) < 5$. In Figure 4.24 the amount of a confounding (γ_{01}) is equal to 5, but the structure of the results is the same also for $\gamma_{01} = 1$, see the additional Figure 42 on page 53 of the digital appendix. The results for all conditions, that means for all simulated datasets with various levels of heterogeneous residual variances, are discussed in section 4.7.

[37] Similar results for unequal group sizes are included as additional figures in the digital appendix: Figure 43 on page 54 and Figure 44 on page 55 for $\gamma_{01} = 1$, and Figure 45 on page 56 and Figure 46 on page 57 for $\gamma_{01} = 5$.

Note that for unequal group sizes and the smallest simulated sample size in part I of the simulation study ($N = 100$), even the simple multi-group model based on the population value of the group size yields inflated empirical type-I-error rates (see the first row in the lower parts of the additional Figure 45 on page 56 and Figure 46 on page 57 of the digital appendix). The observed structure of the rejection frequencies for small sample sizes is different from the observed pattern for larger sample sizes. We will discuss the small sample behavior of the maximum likelihood estimation in subsection 4.8.1.

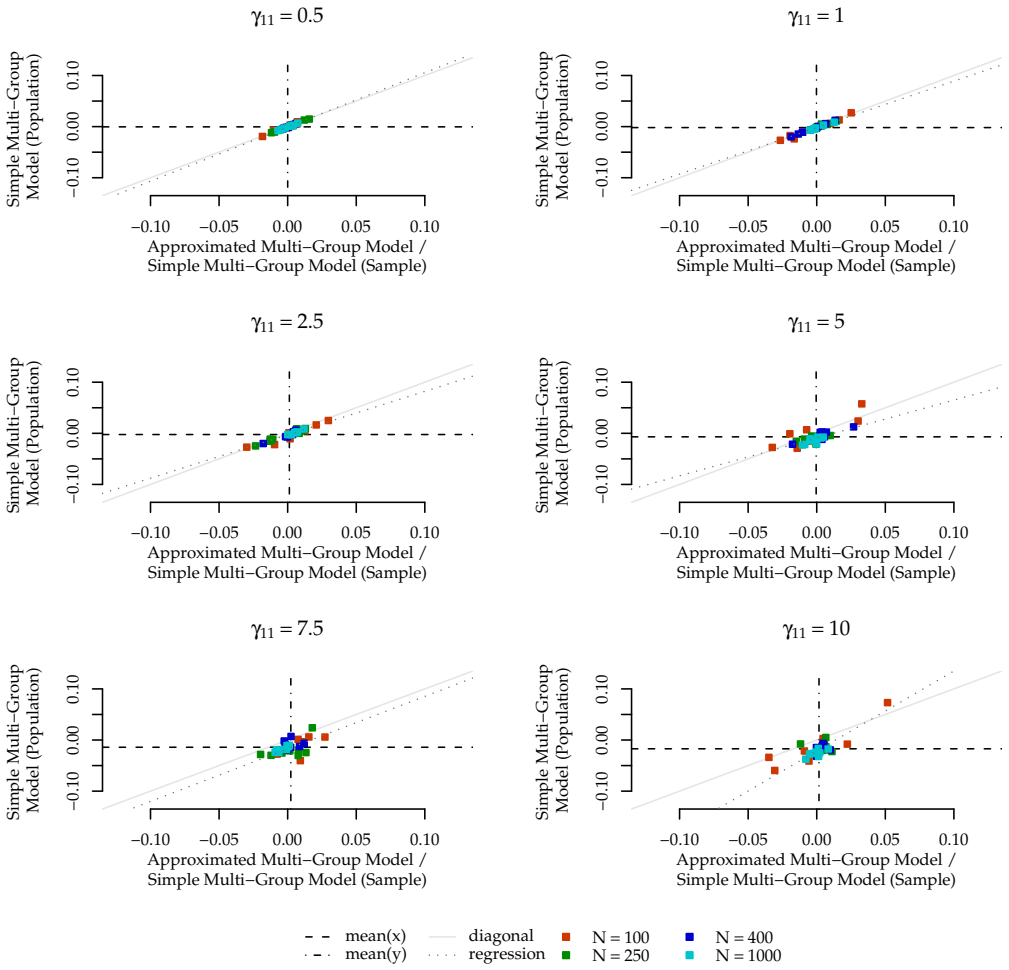


Figure 4.25: Absolute bias of the *ATE*–estimator: Scatter plots for a comparison of the simple multi-group models (population vs. sample), grouped by interaction γ_{11} [$R^2_{X|Z} = 0.75$, $P(X = 1) = 0.5$ and $Var(\varepsilon_{\delta_{10}}) = 0.5$]

Absolute Bias of the *ATE*–Estimator The absolute biases of the *ATE*–estimator for the simple multi-group model based on the true group size (population, *y*–axis), and for the simple multi-group model based on the estimated group size in the constraint (sample, *x*–axis) are plotted in Figure 4.25 for datasets generated with $R^2_{X|Z} = 0.75$, $P(X = 1) = 0.5$ and $Var(\varepsilon_{\delta_{10}}) = 0.5$.[38] For the selected conditions, the simple multi-group model based on the population value of the group size is on average slightly negatively biased for large interaction effects (see the horizontal dotted lines in Figure 4.25).

[38]Note that in Figure 4.25 the observed biases for the multi-group model based on the estimated group size and the approximated multi-group model are combined on the *x*–axis, as the parameter estimates as well as the computation of the point estimate of the average total effect are numerically identical for both procedures.

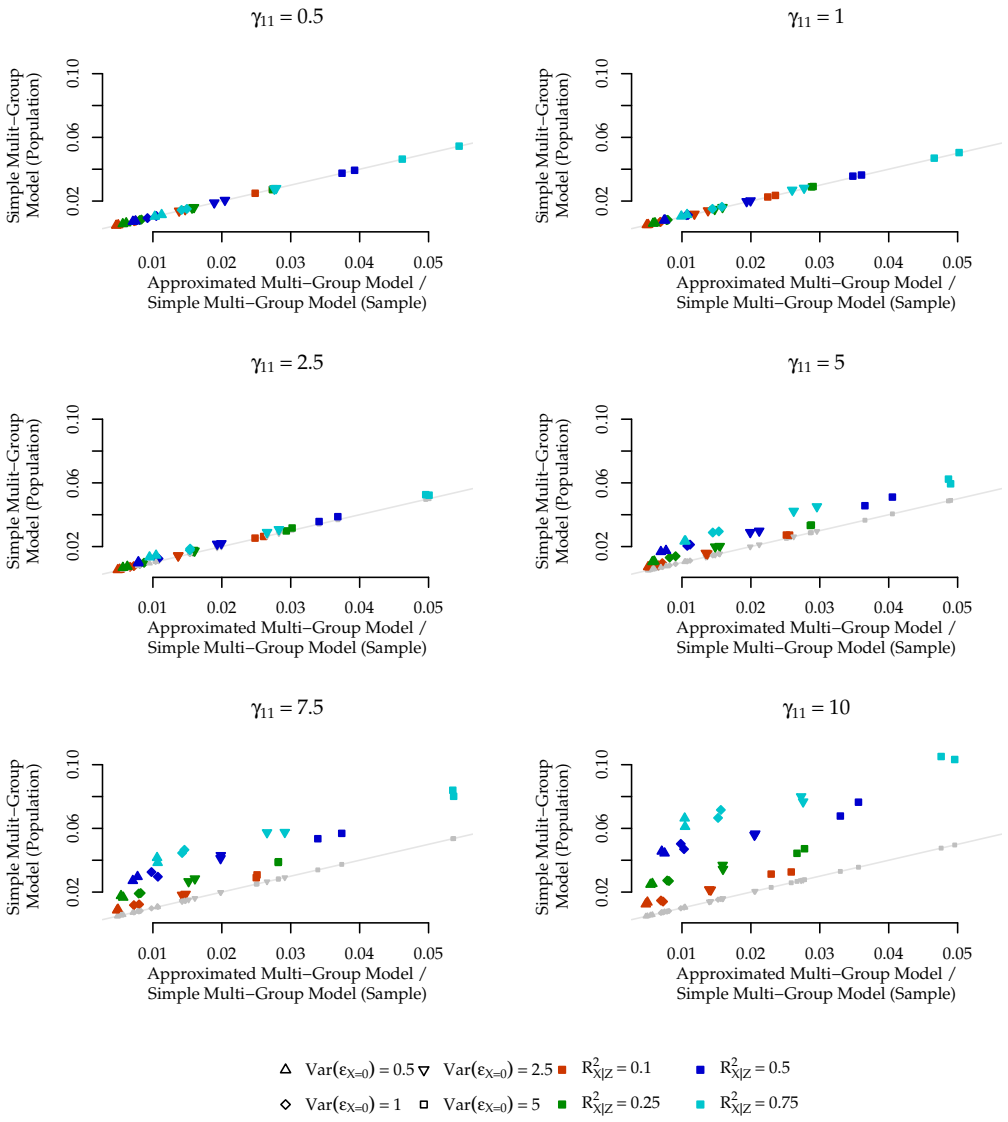
The $B(\widehat{ATE}_{10})$'s are summarized in Table 4.6 for two selected levels of the dependency between X and Z (medium, $R^2_{X|Z} = 0.25$ as well as a strong, $R^2_{X|Z} = 0.75$).[39] This table is similar to table 4.1 in Flory (2008), and most of the computed biases are almost zero. As marked with bold numerics, biases for the average total effect estimator can be observed for the simple multi-group model based on the population value of the group size within some conditions of the simulation study I, particularly for large interaction effects. Nevertheless, the different empirical type-I-error rates for the simple multi-group model based on the population value of the group size versus the simple multi-group model with the sample estimate of the group size in the nonlinear costraint can not be solely explained by $B(\widehat{ATE}_{10})$. The simple multi-group model (population) produces a slightly biased *ATE*–estimator for some conditions under which the average total effect estimator from the simple multi-group model based on the estimated group size is unbiased.

Mean Squared Error of the *ATE*–Estimator Incorporating the true group size for the nonlinear constraint decreases the efficiency of the *ATE*–estimator (as described for the GLH / mean-centering approach above). Furthermore, the absolute biases considered in the last paragraph are computed as the simple sum over the differences between the true ATE_{10} and the estimated $\widehat{ATE}_{10}$, divided by the number of replications (see section 4.4.1). Therefore, positive and negative biases might balance each other out to zero if they are symmetric around the true average total effect. For this reason we also report mean squared errors of the *ATE*–estimator in Figure 4.26 (for equal group sizes)[40] as scatter plots to compare the simple multi-group models (based on the estimated group size on the x–axis vs. based on the true population value of the group size on the y–axis). Obviously, the simple multi-group model based on the population value of the group size yields larger mean squared errors for the *ATE*–estimator (compared to the $MSE\big[\widehat{ATE}_{10}\big]$ obtained for the simple multi-group model constructed with the observed group size in the nonlinear constraint) because random fluctuations of the group size due to sampling error are not taken into account appropriately. For the extreme condition (strong interaction effect, $\gamma_{11} = 10$, and strong dependency of X and Z, $R^2_{X|Z} = 0.75$; see the light blue symbols in the right scatter plot on the bottom of Figure 4.26) the mean squared errors for the *ATE*–estimator are more than twice as large for the population version of the simple multi-group model as compared to the sample version.[41]

[39] See the additional Table 5 on page 125 of the digital appendix for conditions with unequal group sizes. A comparison of Figure 4.25 to the corresponding figures for unequal group sizes, provided as additional figures Figure 47 and Figure 48 on page 58 and 59 of the digital appendix, reveal that the absolute biases are generally larger for conditions with unequal group sizes.

[40] The scatter plots are grouped by the different levels of the interaction parameter γ_{11} used for data generation. The color of the symbols indicates the levels of the relationship between X and Z (varied in the data generation as $R^2_{X|Z}$) within each of the six charts. Furthermore, different symbols are used to refer to the amount of residual variance $Var(\varepsilon_{X=0})$. The additional Figure 49 on page 60 and Figure 50 on page 61 of the digital appendix present the same conditions for unequal group sizes.

[41] Identical mean squared errors are obtained for the simple multi-group model based on the sample value of the group size and for the approximated multi-group model (see the small gray symbols in the scatter plots, which are all on the diagonal line). Therefore, the approximated multi-group model was added as an additional label for the description of the x–axis.



Note: Small gray symbols represent the results of the simple multi-group model (sample) on the x–axis and the approximated multi-group model on the y–axis.

Figure 4.26: Mean squared error of the *ATE*–estimator: Scatter plots for a comparison of the simple multi-group models (sample vs. population), grouped by interaction γ_{11} [$N = 1000$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $P(X = 1) = 0.5$]

Different patterns of the $MSE[\widehat{ATE}_{10}]$ for the two versions of the simple multi-group model can also be distinguished in Figure 4.27. The *ATE*–estimators $MSE[\widehat{ATE}_{10}]$ for both simple multi-group models decrease with increasing sample size and increase with increasing dependency of X and Z, as well as with increasing residual variance $Var(\varepsilon_{X=0})$ [for the selected conditions of the simulation study presented in

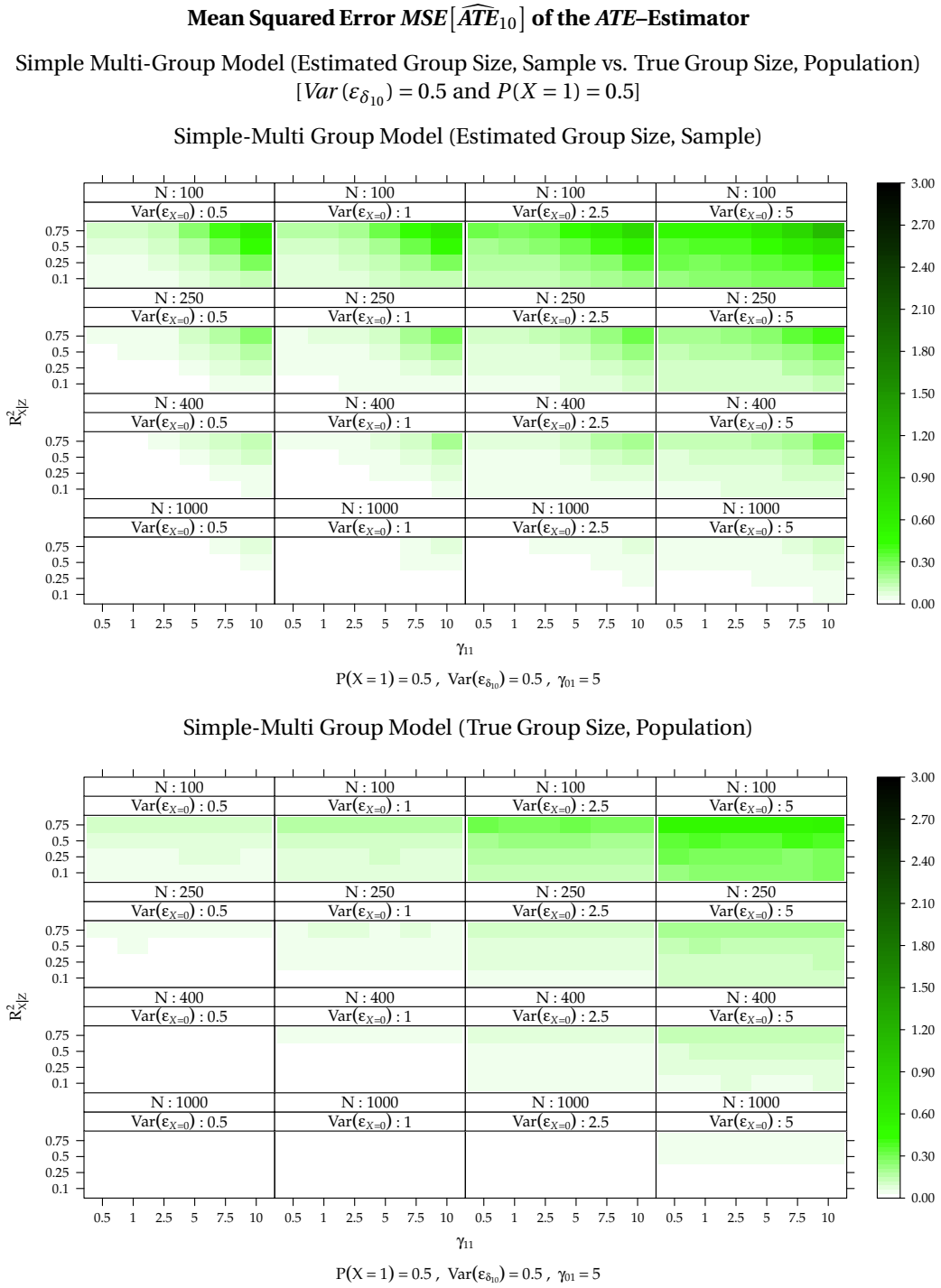


Figure 4.27: Mean squared error of the *ATE*–estimator: Level plots for a comparison of the simple multi-group models (sample vs. population) [$Var(\varepsilon_{\delta_{10}}) = 0.5$ and $P(X = 1) = 0.5$]

Figure 4.27]. Despite these similarities, the $MSE\left[\widehat{ATE}_{10}\right]$ increases only for the simple multi-group model based on the estimated group size along with the amount of interaction (γ_{11}) used for generating the data.[42]

[42]Note that this result is less obvious for unequal group sizes, see the additional figures 46 and 45 on page 57 and 56 of the digital appendix.

Table 4.7: Relative bias of the standard error of the *ATE*–estimator: Comparison of the simple multi-group models (population vs. sample) [$P(X = 1) = 0.5$, $\gamma_{01} = 5$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $Var(\varepsilon_{X=0}) = 0.5$]

		Interaction (γ_{11})					
$R^2_{X\|Z}$	N	0.5	1.0	2.5	5.0	7.5	10.0
Simple Multi-Group Model (Population)							
0.1	100	4.90 %	4.50 %	−4.81 %	−1.59 %	−1.89 %	−1.13 %
	250	−1.42 %	−0.68 %	0.70 %	3.22 %	0.59 %	−1.34 %
	400	−1.74 %	2.46 %	2.61 %	3.54 %	1.37 %	−3.93 %
	1000	1.95 %	−1.22 %	−1.89 %	−2.86 %	−3.09 %	−1.27 %
0.25	100	−0.93 %	0.09 %	−0.28 %	2.45 %	−0.97 %	−0.91 %
	250	−4.14 %	1.86 %	0.39 %	1.93 %	0.33 %	2.28 %
	400	−0.36 %	2.09 %	−0.66 %	−1.67 %	0.74 %	−2.35 %
	1000	−2.07 %	−0.72 %	−3.53 %	−1.32 %	−0.47 %	2.75 %
0.5	100	−1.20 %	−1.05 %	−1.42 %	−2.53 %	−4.18 %	1.20 %
	250	1.32 %	−3.25 %	2.27 %	2.48 %	−1.81 %	3.08 %
	400	−0.21 %	1.83 %	−1.61 %	−1.27 %	−4.97 %	2.13 %
	1000	2.41 %	−1.99 %	−3.24 %	−2.55 %	3.30 %	−3.19 %
0.75	100	−0.79 %	−0.07 %	−3.75 %	0.47 %	0.16 %	−4.99 %
	250	−1.46 %	−3.77 %	1.61 %	1.27 %	2.59 %	−0.80 %
	400	2.57 %	−0.95 %	1.17 %	−0.01 %	−1.72 %	−1.05 %
	1000	−2.79 %	−2.02 %	1.23 %	−2.89 %	1.44 %	−1.06 %
Simple Multi-Group Model (Sample)							
0.1	100	4.21 %	3.83 %	**-7.04** %	**-5.37** %	**-6.55** %	**-5.36** %
	250	−1.46 %	−0.90 %	−1.33 %	0.47 %	−3.96 %	−4.70 %
	400	−2.21 %	1.44 %	−0.23 %	0.16 %	−3.65 %	**-8.05** %
	1000	1.59 %	−1.18 %	−4.08 %	**-6.64** %	**-7.01** %	−4.62 %
0.25	100	−1.24 %	−0.20 %	**-5.48** %	**-5.90** %	**-8.75** %	**-8.18** %
	250	−4.41 %	0.56 %	−4.84 %	**-5.38** %	**-9.49** %	**-7.38** %
	400	−1.03 %	0.20 %	−3.80 %	**-8.53** %	**-8.62** %	**-10.91** %
	1000	−2.25 %	−2.52 %	**-8.13** %	**-9.83** %	**-9.23** %	**-6.67** %
0.5	100	−1.70 %	−3.15 %	**-11.37** %	**-16.29** %	**-19.86** %	**-19.35** %
	250	0.59 %	−4.04 %	−3.76 %	**-13.76** %	**-16.65** %	**-17.88** %
	400	−0.93 %	−2.18 %	**-10.14** %	**-15.73** %	**-20.78** %	**-19.33** %
	1000	1.66 %	−3.41 %	**-12.65** %	**-14.95** %	**-16.95** %	**-20.28** %
0.75	100	−1.46 %	−2.93 %	**-12.61** %	**-21.77** %	**-26.17** %	**-29.68** %
	250	−2.28 %	**-6.23** %	**-9.13** %	**-22.49** %	**-25.76** %	**-29.44** %
	400	2.01 %	−3.10 %	**-8.81** %	**-20.77** %	**-28.14** %	**-26.79** %
	1000	−3.29 %	−2.94 %	**-9.37** %	**-21.52** %	**-22.45** %	**-29.65** %

Bias of the Standard Error of the *ATE*–Estimator To reconcile our results with the simulation study conducted by Flory (2008), we finally present the *ATE*–estimators' relative bias of the standard error for both simple multi-group models. For the relative biases, Flory considered absolute values larger than five percent as severely biased. Following the same rule, Table 4.7 presents the results obtained in the first part of our simulation study. The upper part of the table shows the relative biases of the *ATE*–estimators' standard error for the simple multi-group model based on the true population value of the group size, and the lower part of the table shows the same dependent measure tabulated for the model based on the estimated group size. Obviously, the simple multi-group model based on the sample estimate of the group size yields biased

standard errors for the average total effect estimator in almost all conditions with a substantial dependency between X and Z, medium to large interaction effects, and with equal group sizes. Similar findings were observed for conditions with unequal group sizes,[43] even though the relative biases of the average total effect estimators' standard error are slightly smaller, especially for $P(X = 1) = 0.8$.

Convergence Rate It should be noted that the convergence rates of both simple multi-group models discussed in this section as well as of all other multi-group models considered in this thesis (the elaborated multi-group model and the approximated multi-group model) are generally inconspicuous for the studied sample sizes in simulation study I: All multi-group models converged in almost all conditions, i. e., the observed convergence rates are >= 99.99 %.

Summary The results presented in this subsection for the first part of the simulation study support the conclusion that the simple multi-group model based on the sample estimate of the group size yields an unbiased estimator of the average total effect. If the nonlinear constraint for the hypothesis $ATE = 0$ is computed with the help of the sample estimate of the group size, the estimated variance of the *ATE*–estimator does not appropriately account for the stochasticity of the treatment variable. This results in an underestimation of the asymptotic variance of the *ATE*–estimator and yields subsequently inflated empirical type-I-error rates. The amount of the underestimation of the average total effect estimators' standard error depends on the amount of interaction (γ_{11}) and is a function of the dependency of X and Z ($R^2_{X|Z}$). Consequently, the simple multi-group model cannot be suggested as a trustable implementation of generalized analysis of covariance.

4.6.2 Simple Single Group Model

In addition to the simple multi-group models discussed in the previous subsection, Flory (2008) suggested a single group structural equation model as a possible alternative implementation of generalized analysis of covariance. We introduced this *simple single group model (with interaction)* in section 3.3.4.1 in detail and pointed out that this model is expected to perform well for all conditions of the simulation design, where datasets were generated with a small amount of residual variance heterogeneity. The performance of the simple single group model under these conditions of the first part of the Monte Carlo simulation is discussed in this subsection. Furthermore, we will return to the results obtained for the simple single group model for conditions with significant between-group residual variances heterogeneity in section 4.7, where we compare this model to the more appropriate generalization, the *elaborated single group model* as developed in subsection 3.3.4.2.

[43] See the additional Table 6 on page 126 and Table 7 on page 127 of the digital appendix for the corresponding results for unequal group sizes. Furthermore, the additional Figure 53 – 55 present the overall structure of the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$, see pages 64 – 66 of the digital appendix.

Type-I-Error Rate The simple single group model performed, as expected, reasonably well for equal group sizes and almost homogeneous residual variances, i. e., we were able to replicate the findings presented by Flory (2008) for datasets generated under similar conditions. The observed rejection frequencies are presented in Figure 4.28.

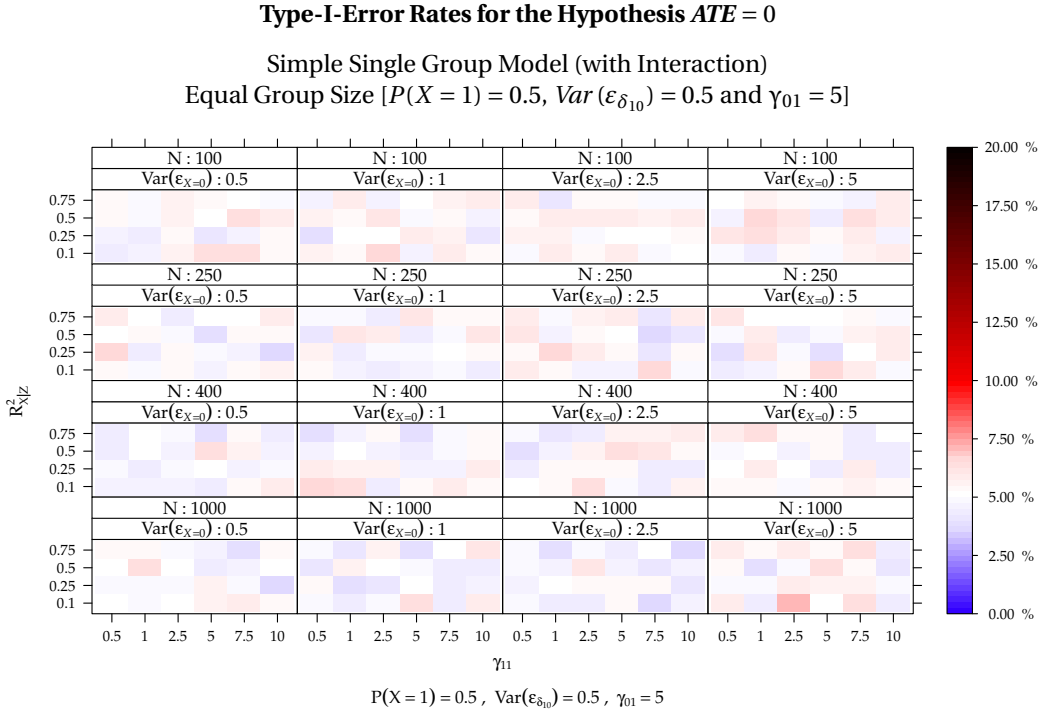


Figure 4.28: Type-I-error rate: Level plot for the simple single group model (with interaction) [$P(X = 1) = 0.5$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $\gamma_{01} = 5$]

Unfortunately, this finding is not generalizable to conditions with unequal group sizes and almost homogeneous residual variances,[44] as shown in Figure 4.29. The upper part of this figure shows the empirical type-I-error rates for a treatment group smaller than its control group [$P(X = 1) = 0.2$], while the lower part shows the empirical type-I-error rates for conditions with a treatment group larger than its control group [$P(X = 1) = 0.8$]. The first and the second columns in Figure 4.29 present the observed h_{RF}'s of the simple single group model under conditions with mild violations of the implied variance structure [$Var(\varepsilon_{X=0}) < 2.5$ and $Var(\varepsilon_{\delta_{10}}) = 0.5$].[45] Obviously, even under these conditions and for small interaction effects ($\gamma_{11} < 5$) the

[44] For the interpretation of the pattern in Figure 4.29 it is important to remember the exact procedure we used for generating the data (see section 4.2 for details). The amount of the violation of the homogeneity assumption depends on the following two parameters: The residual variance $Var(\varepsilon_{X=0})$ [i. e., the residual variance of the covariate-(treatment) regression in the control group] and the residual variance $Var(\varepsilon_{\delta_{10}})$ for the regression of the individual total effect on the covariate. For theoretical reasons, no conditions with $Var(\varepsilon_{\delta_{10}}) = 0$ were included in the simulation study. Hence, for the subset of simulated conditions with the smallest unexplained variance of the individual total effect [$Var(\varepsilon_{\delta_{10}}) = 0.5$] presented in this section as *condition where homogeneity of residual variance almost holds*, the degree of the violation of the homogeneity assumption depends on the value of $Var(\varepsilon_{X=0})$ [varying from 0.5 being a strong violation of the assumption of residual variance heterogeneity to 5, the smallest amount of residual variance heterogeneity considered in simulation study I].

[45] Conditions with strong heterogeneity of residual variance will be discussed in the subsequent section 4.7.

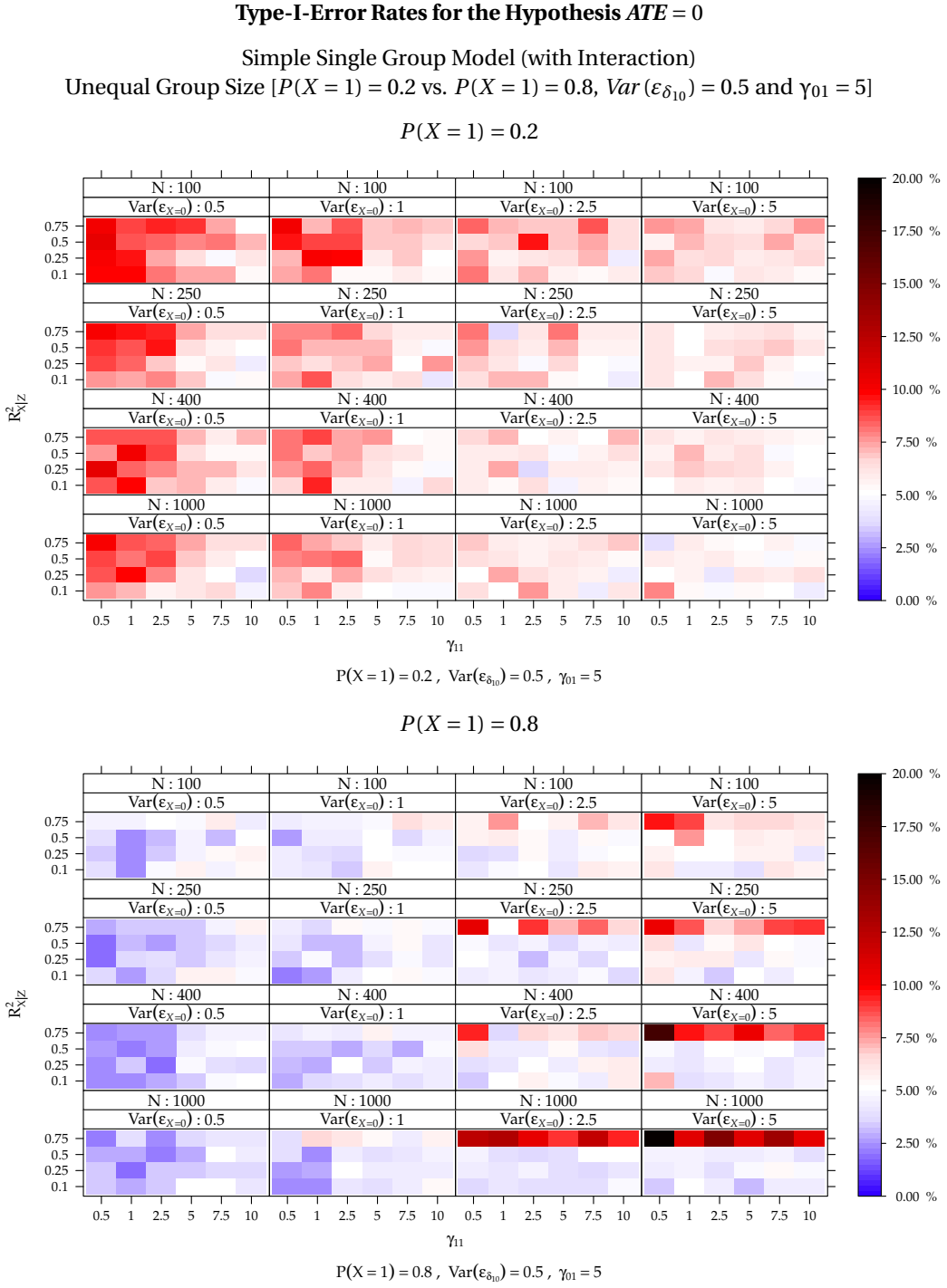


Figure 4.29: Type-I-error rate: Level plot for the simple single group model (with interaction) [$P(X = 1) = 0.2$ and $P(X = 1) = 0.8$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $\gamma_{01} = 5$]

empirical type-I-error rates are inflated for unequal group sizes with $P(X = 1) = 0.2$. Consequently, the empirical type-I-error rates fall below the nominal level for the same conditions of the simulation study for unequal group sizes with $P(X = 1) = 0.8$. Both effects (too high and too low empirical type-I-error rates) are

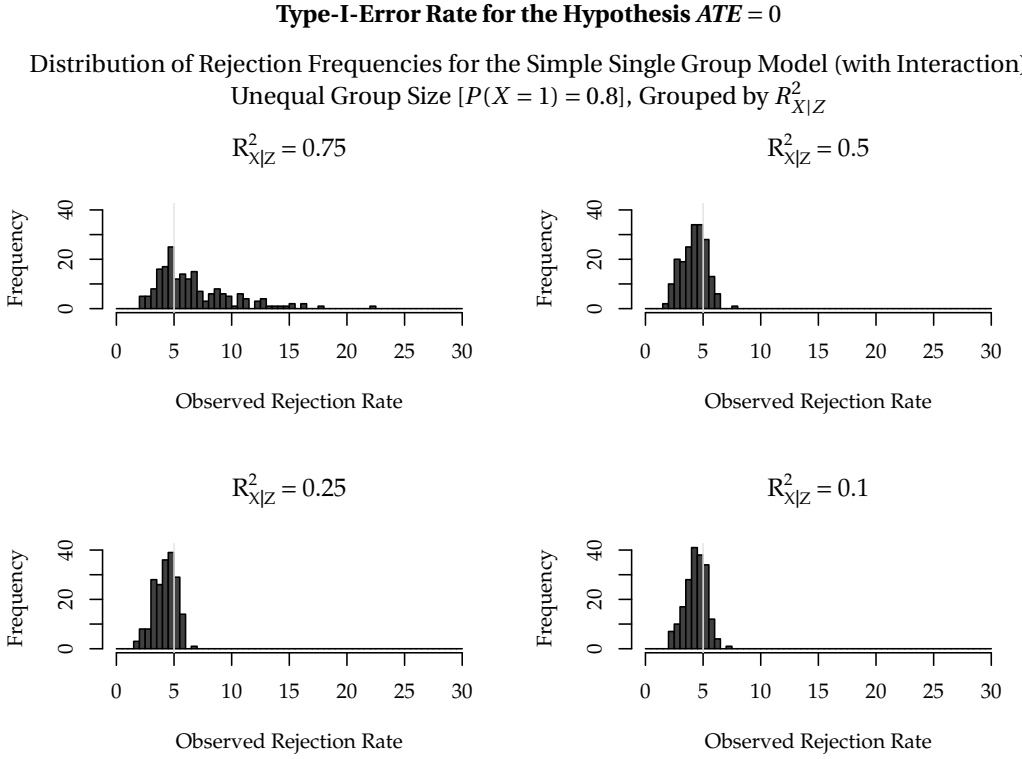


Figure 4.30: Type-I-error rate: Histograms for the distribution of observed rejection frequencies for the simple single group model (with interaction), grouped by $R^2_{X|Z}$ [$P(X = 1) = 0.8$ and $Var(\varepsilon_{\delta_{10}}) = 0.5$]

connected to the interaction paramter γ_{11} because the nominal type-I-error rate for very srong covariate-treatment interactions ($\gamma_{11} = 10$) was achieved in most of the presented conditions.[46]

A surprising exception of the observed pattern of rejection frequencies for the simple single group model is visible in the lower part of Figure 4.29: The empirical type-I-error rates are heavily inflated for conditions with a treatment probability of $P(X = 1) = 0.8$ and a strong dependency between X and Z ($R^2_{X|Z} = 0.75$) for conditions with only minor violations of the homogeneity assumption [$Var(\varepsilon_{X=0} > 1$]. These abnormalities are most obvious for the largest studied sample size in the simulation study I ($N = 1000$). The empirical distributions of rejection frequencies (approximated as histograms) for the simple single group model for all conditions with $Var(\varepsilon_{\delta_{10}}) = 0.5$ — grouped by $R^2_{X|Z}$ — give a detailed view on this finding (see Figure 4.30). The observed distribution of the rejection frequencies changes dramatically from negative to positive skewed in the transition from $R^2_{X|Z} = 0.5$ to $R^2_{X|Z} = 0.75$.[47]

[46] Although the empirical type-I-error rate is reported here, almost the same systematic pattern of results can be observed by inspecting the relative bias of the *ATE*–estimators standard error, presented in the additional Figure 56 on page 67, Figure 57 on page 68 and Figure 58 on page 69 of the digital appendix.

[47] It is interesting to note that the same effect is not that clearly visible for unequal treatment groups with $P(X = 1) = 0.2$ (see Figure 59 on page 70 of the digital appendix).

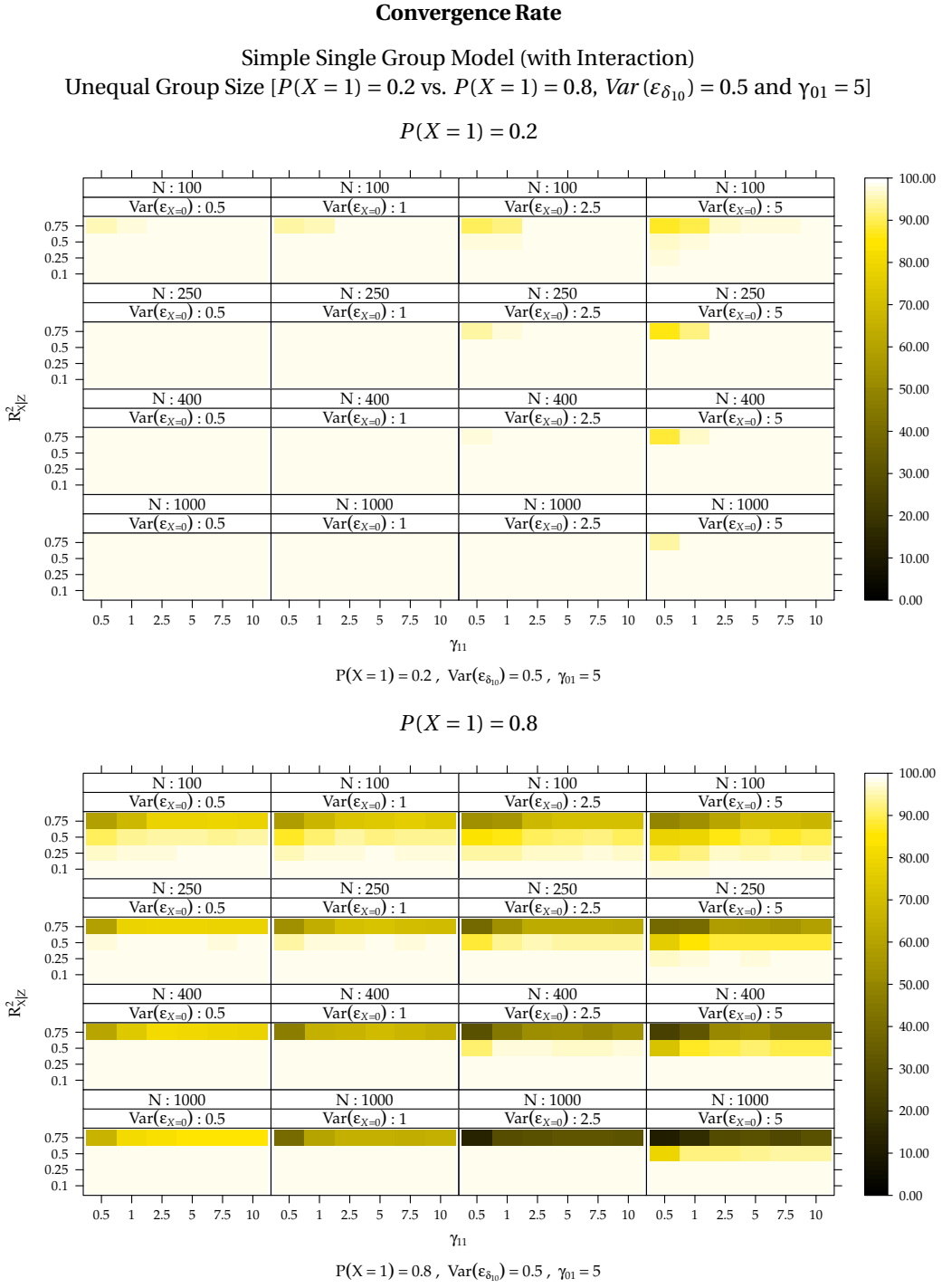


Figure 4.31: Convergence rate: Level plots for the simple single group model (with interaction) [$P(X = 1) = 0.2$ and $P(X = 1) = 0.8$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $\gamma_{01} = 5$]

Convergence Rate The unexpected inflated empirical type-I-error rates for conditions with strong dependencies between X and Z and unequal group sizes can be explained by the observed convergence rates of

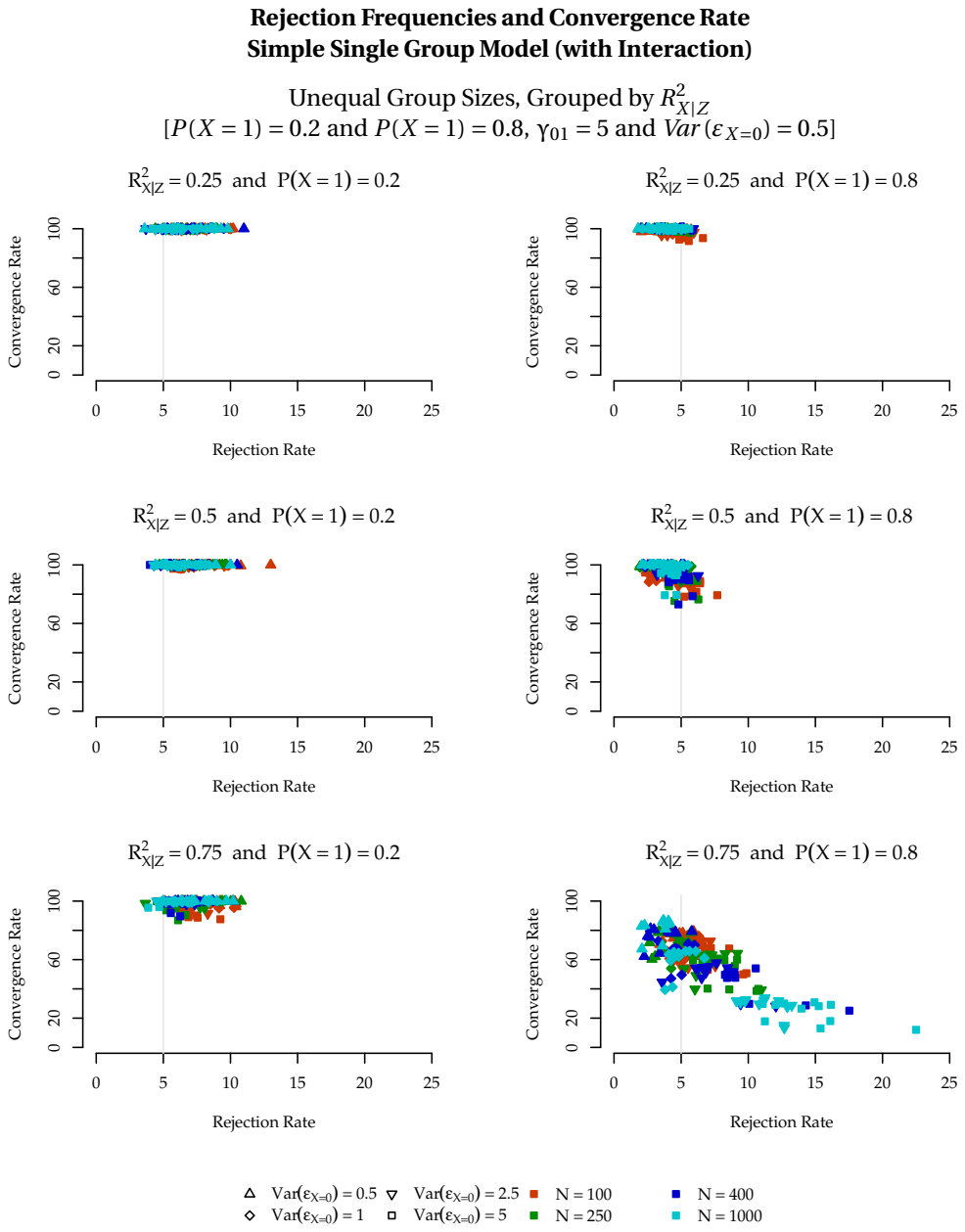


Figure 4.32: Rejection frequencies and the convergence rates for the simple single group model (with interaction), grouped by $R^2_{X|Z}$ [$P(X=1)=0.2$ and $P(X=1)=0.8$, $\gamma_{01}=5$ and $Var(\varepsilon_{X=0})=0.5$]

the simple single group model.[48] The simple single group models' substantive convergence problems are exemplified in Figure 4.31. For all conditions with $R^2_{Z|X} > 0.5$ and especially for the conditions with treatment groups larger than the control groups (compare upper and lower parts of Figure 4.31), the estimation of the `Mplus`–model converged only infrequently. Obviously, the models' convergence rates decrease as the value of $R^2_{X|Z}$ increases, and the convergence rates increase with the amount of interaction (γ_{11}) within

[48] As noted in section 4.4.5, rejection frequencies are computed based on the number of converged replications within each cell of the simulation studies design.

each level of the dependency between X and Z. Overall, an additional effect of the sample size on the convergence rates is observable: For treatment groups larger than the control groups the convergence rates are higher for larger sample sizes. For treatment groups smaller than the control group the convergence rates are high for small dependencies between X and Z and decrease for strong dependencies (compare the rows in the lower part of Figure 4.31).[49]

Finally, for the simple single group model (with interaction) we observed a relationship between the rejection frequencies and the convergence rates which depends systematically on the dependency of X and Z. This phenomenon is displayed in Figure 4.32. Additinaly, within some parts of the simulation studies design (i. e., for a given value of the treatment probability and a fixed dependency of X and Z), the rejection frequencies are related to the sample size (the different sample size conditions are represented with different colors in Figure 4.32).

Absolute Bias of the *ATE*–Estimator Although the Wald–test of the hypothesis $ATE = 0$ based on the simple single group model (with interaction) does not adhere to the nominal α–level (because the model is not robust to minor violations of the implied variance structure), we found that the average total effect estimator is unbiased for equal group sizes under (almost) homogenous residual variances (see Figure 4.33).

This observation confirms the findings presented by Flory (2008) and is useful for later comparisons. Only under a few selected single conditions of the simulation studies design [i. e., for large interaction parameters γ_{11}, small sample sizes $N = 100$ and large residual variances $Var(\varepsilon_{X=0})$] do the obtained absolute biases of the *ATE*–estimator depart vertically from the diagonal lines in Figure 4.33. On average (marked as dotted lines) the simple single group model and the elaborated single group model give unbiased average total effect estimators for equal group sizes and homogeneity of residual variance.

The $B(\widehat{ATE}_{10})$ is obviously not always zero for conditions of the simulation study with unequal group sizes (see Figure 4.34). The simple single group model yields a biased *ATE*–estimator for simulated datasets with small interaction effects, in particual for large residual variances $Var(\varepsilon_{X=0}) = 5$ (represented in light blue in Figure 4.34). A systematic negative bias is observable for $P(X = 1) = 0.8$ and $\gamma_{11} < 5$ [which is distinct from the $B(\widehat{ATE}_{10})$ of the more appropriate elaborated single group model, see the discussion in subsection 4.7.1].[50]

[49] We observed no convergence problems for the simple single group model for conditions with equal group sizes. Detailed plots of the convergence rates for all studied methods and for all simulated conditions are given in the digital supplement (see Digital Supplement: 1-6).

[50] As shown in the additional Figure 61 on page 72 of the appendix, we also observed a biased *ATE*–estimator for the elaborated single group model for $P(X = 1) = 0.2$. This bias of the *ATE*–estimator will be discussed in more detail in subsection 4.7.1. Furthermore, the additional Table 8 on page 128 of the digital appendix presents the absolute biases for conditions of the simulation study with unequal group sizes, $\gamma_{01} = 5$, $Var(\varepsilon_{\delta_{10}}) = 0.5$ and $Var(\varepsilon_{X=0}) = 5$.

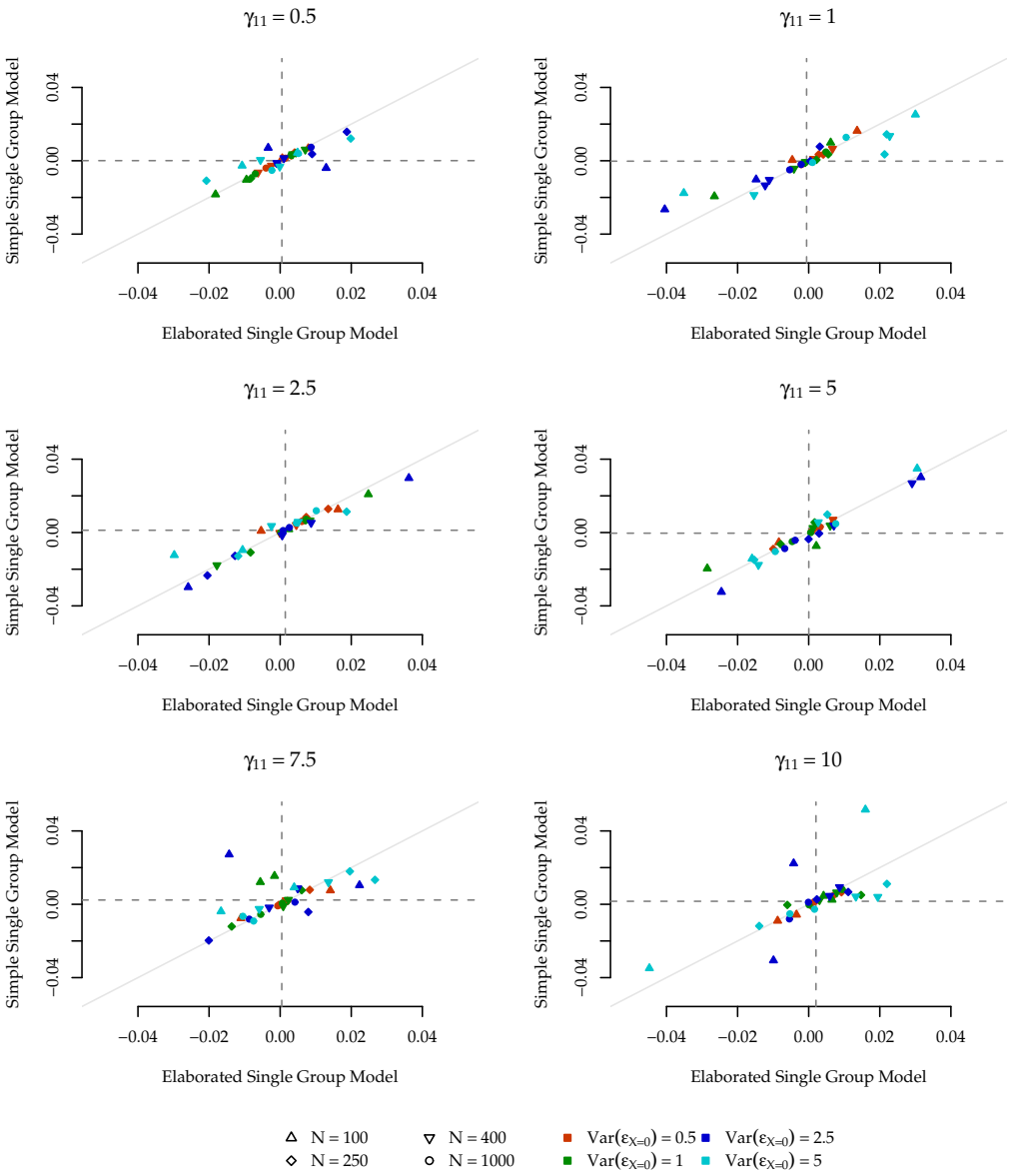


Figure 4.33: Absolute bias of *ATE*–estimator: Scatter plots for a comparison of the simple single group model (with interaction) and the elaborated single group model, grouped by interaction γ_{11} $[P(X=1)=0.5, R^2_{X|Z}=0.75$ and $Var(\varepsilon_{\delta_{10}})=0.5]$

Summary By means of the Monte Carlo simulation we found that, analogous to the methods based on the general linear model, the relative group size is of major importance for the robustness of the single group structural equation models: For equal group sizes, the simple single group model (with interaction) offers a feasible way for implementing generalized analysis of covariance within the framework of structural equa-

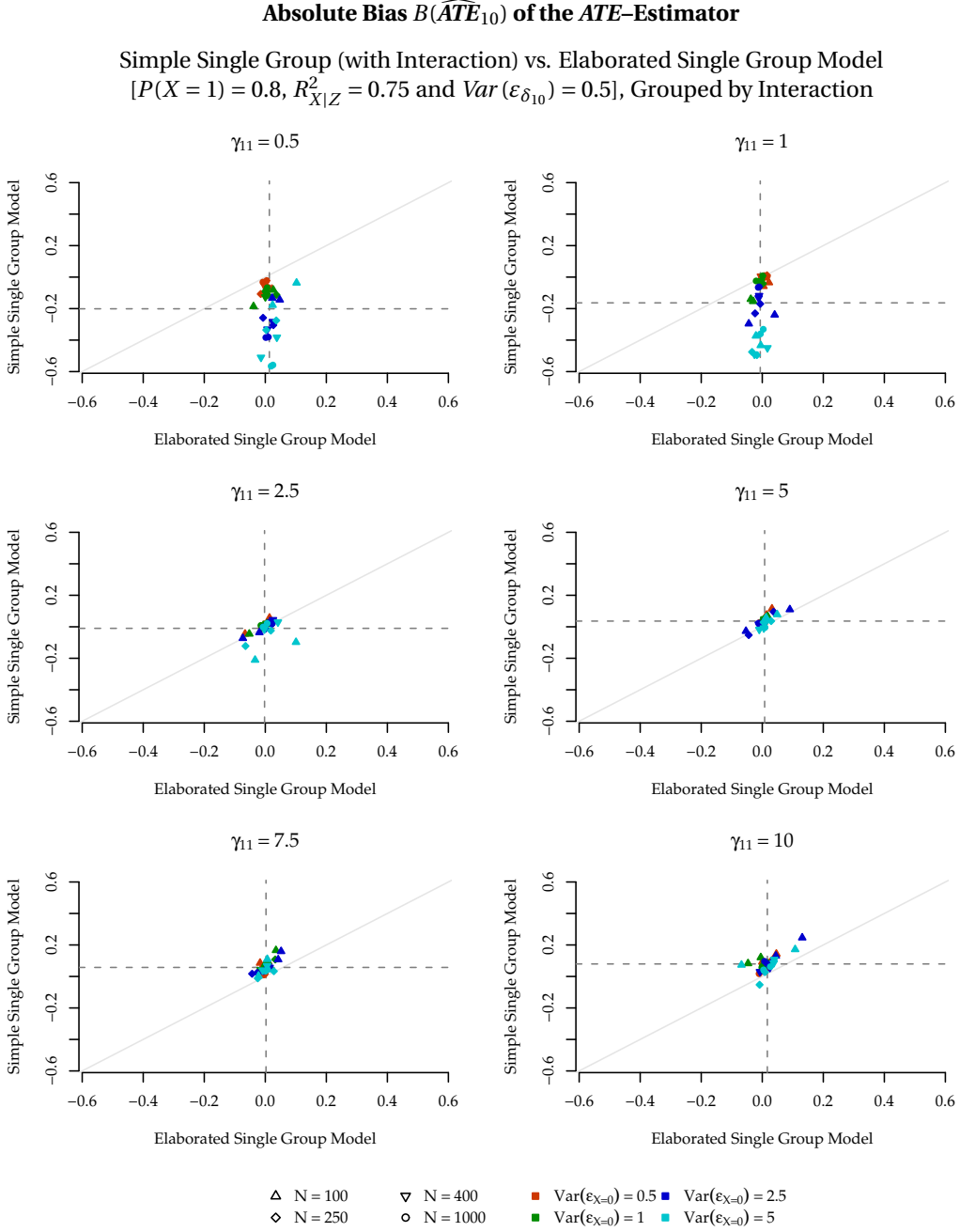


Figure 4.34: Absolute bias of *ATE*–estimator: Scatter plots for a comparison of the simple single group model (with interaction) and the elaborated single group model, grouped by interaction γ_{11} [$P(X=1)=0.8$, $R^2_{X|Z}=0.75$ and $Var(\varepsilon_{\delta_{10}})=0.5$]

tion modeling. This conclusion cannot be generalized to conditions with unequal group sizes. The test of the hypothesis of no average total effect based on this model yields empirical type-I-error rates above (for treatment groups smaller than the control group) or below (for treatment groups larger than the control group) the nominal level even under conditions which nearly fulfill the assumption of homogeneous resid-

ual variances. Two different problems explain the missing robustness of the simple single group model: The notable bias of the *ATE*–estimator's standard error for unequal group sizes and serious convergence problems observed for $P(X = 1) = 0.8$. As a consequence of the misspecified implied variance structure we observed the *ATE*–estimator to be absolutely biased for unequal group siezs. Hence, an implementation of generalized analysis of covariance based on the simple single group model seems unwarrantable. [51]

4.6.3 Summary

In this section the results of two implementations of generalized analysis of covariance within the framework of structural equation modeling were presented. Due to the maximum likelihood estimation (see section 3.3.1) both models are estimated under a multivariate normality assumption. This was formulated as a necessary requirement in order to derive valid unconditional inference about the estimated average total effect. Nevertheless, both models were found to be infeasible under the considered conditions.

The simple multi-group model is estimated under a multivariate normality assumption conditional on X. Accordingly, the group size (i. e., the mean of the treatment variable) is not incorporated as an estimated parameter for the nonlinear constraint. We demonstrated that the resulting standard error for the average total effect estimator underestimates the (true) variability of the estimated average total effects.

The simple single group model (with interaction) was shown to violate the formulated requirements for the implementation of generalized analysis of covariance with respect to the implied variances (i. e., equal between-group residual variances are assumed for the model specification, see subsection 3.3.4.1). The results of the Monte Carlo simulation confirmed that the misspecified implied variance structure can lead to biased parameter estimates. To our surprise the simple single group model was obviously not robust against even small violations of the implied variance structure. Furthermore, we observed unexpected convergence problems for this single group structural equation modeling approach.

Therefore, we shall skip a detailed presentation of the *ATE*–estimators standard error under almost homogenous residual variances in this subsection. In the following subsection we will discuss the simple single group model (with interaction) in comparison to the elaborated single group model under all conditions of simulation study I.

[51] The distributions of the rejection frequencies for the test of the hypothesis $ATE = 0$ obtained from the simple multi-group models and from the simple single group model are presented in the digital appendix as additional Figure 62 on page 73.

4.7 Results of Structural Equation Models under Heterogeneity of Residual Variances

4.7.1 Elaborated Single Group Model

In this section we will analyze the performance of the *elaborated single group model* as an extension of the simple single group model under conditions of simulation study I with heterogeneity of between-group residual variances. Overall, the *ATE*–estimator based on this extended single group model was expected to be unbiased, with correct standard errors and consequently nominal type-I-error rates of 5 %. According to the research questions formulated in section 3.4, we will focus on the accuracy of the estimated asymptotic variances and covariances of parameter estimates and especially on the accuracy of the derived standard error of the *ATE*–estimator. In particular, we shall focus on conditions with unequal group sizes and heterogeneity of residual variance, i. e., we shall study the conditions where the simple single group model (with interaction) failed. This is motivated by the previously described finding that the single group structural equation modeling approach is sensitive to a misspecified implied variance structure. Furthermore, we will report the results of both models for conditions of the simulation study where substantial empirical differences were observed.

Absolute Bias of the *ATE*–Estimator The *ATE*–estimator obtained from the simple single group model (with interaction) was found to be biased for some of the simulated conditions with small violations of the implied variance structure (see subsection 4.6.2). Accordingly, we shall continue the comparison of the simple single group model with the elaborated single group model for data generated with a substantial amount of residual variance for the individual total effect [i. e., $Var(\varepsilon_{\delta_{10}})$] in Figure 4.35.

Obviously, the observed bias of the *ATE*–estimator based on the simple single group model for unequal group sizes vanishes by specifying the slope in the regression Y on X as *random slope* (see subsection 3.3.4.2 for details), at least for the printed conditions with the largest sample size ($N = 1000$). Although the *ATE*–estimator is still biased for unequal group sizes and small sample sizes,[52] the systematic patterns we observed for the simple single group model (with interaction) for conditions with unequal group sizes [$P(X = 1) = 0.8$] as a function of the interaction parameter γ_{11} disappear. The elaborated single group model performs much better compared to results obtained for the simple single group model for conditions with strong heterogeneity of residual variance, treatment groups larger than control groups, and small values of the interaction parameter γ_{11}.

[52] See the additional Figure 63 on page 74 of the digital appendix for a comparison of $B(\widehat{ATE}_{10})$ between the simple single group model and the elaborated single group model for $N = 100$.

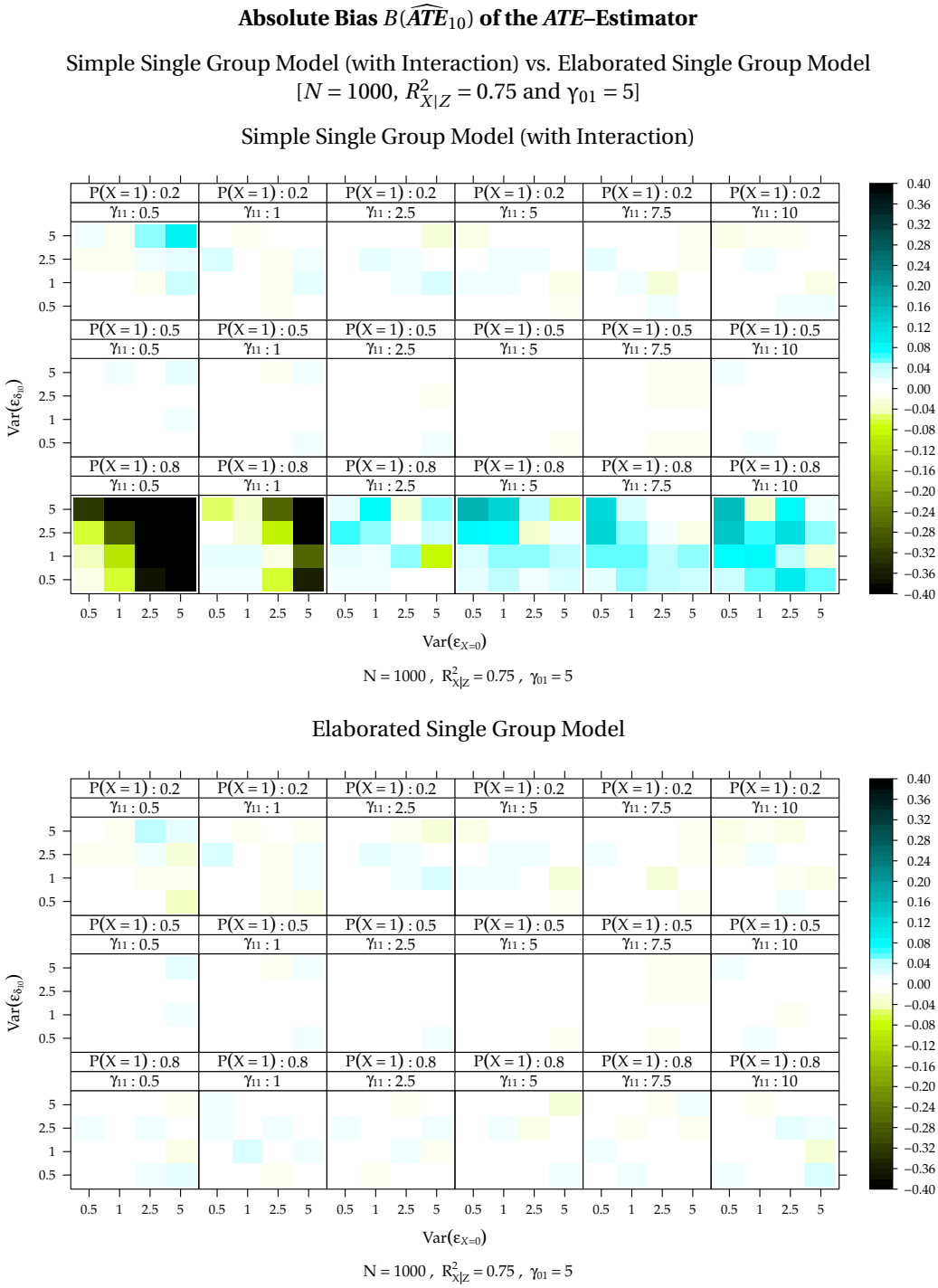


Figure 4.35: Absolute bias of the *ATE*–estimator: Level plots for a comparison of the simple single group model (with interaction) and the elaborated single group model [$N = 1000$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

Figure 4.36 and Figure 4.37 explore a comparison of the elaborated single group model and the simple single group model with respect to $B(\widehat{ATE}_{10})$ under conditions of strong variance heterogeneity further. For

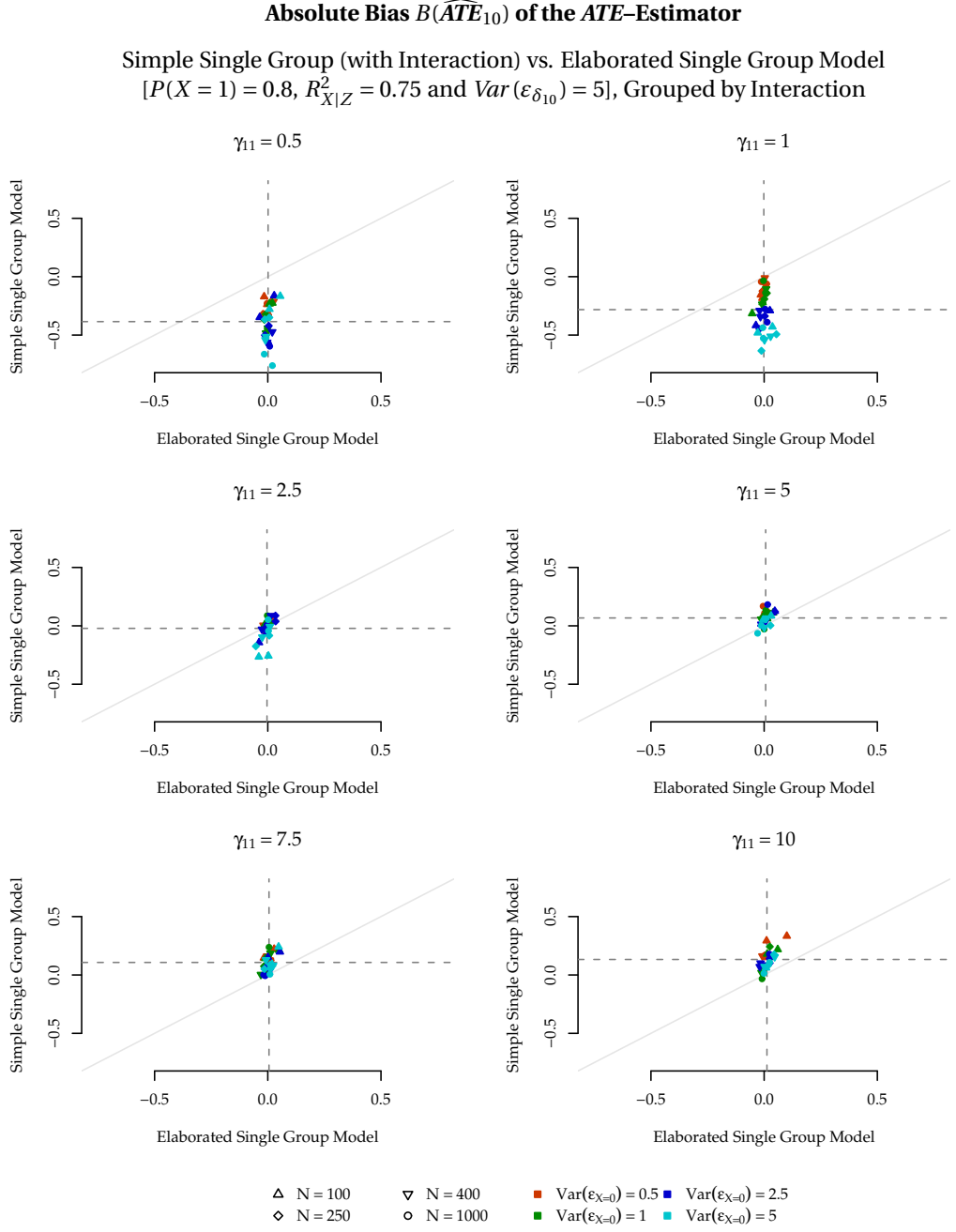


Figure 4.36: Absolute bias of the *ATE*–estimator: Scatter plots for a comparison of the simple single group model (with interaction) and the elaborated single group model, grouped by γ_{11} [$R^2_{X|Z}=0.75$, $P(X=1)=0.8$ and $Var(\varepsilon_{\delta_{10}})=5$]

unequal group sizes — most prominent for $P(X=1)=0.8$ — the simple single group model (with interaction) is not robust at all against the misspecified implied variance structure. Again, a systematic bias of the *ATE*–estimator is prominent (see horizontal dotted lines in Figure 4.36 indicating the absolute bias of *ATE*–estimator obtained from the simple single group model averaged over all conditions with the given

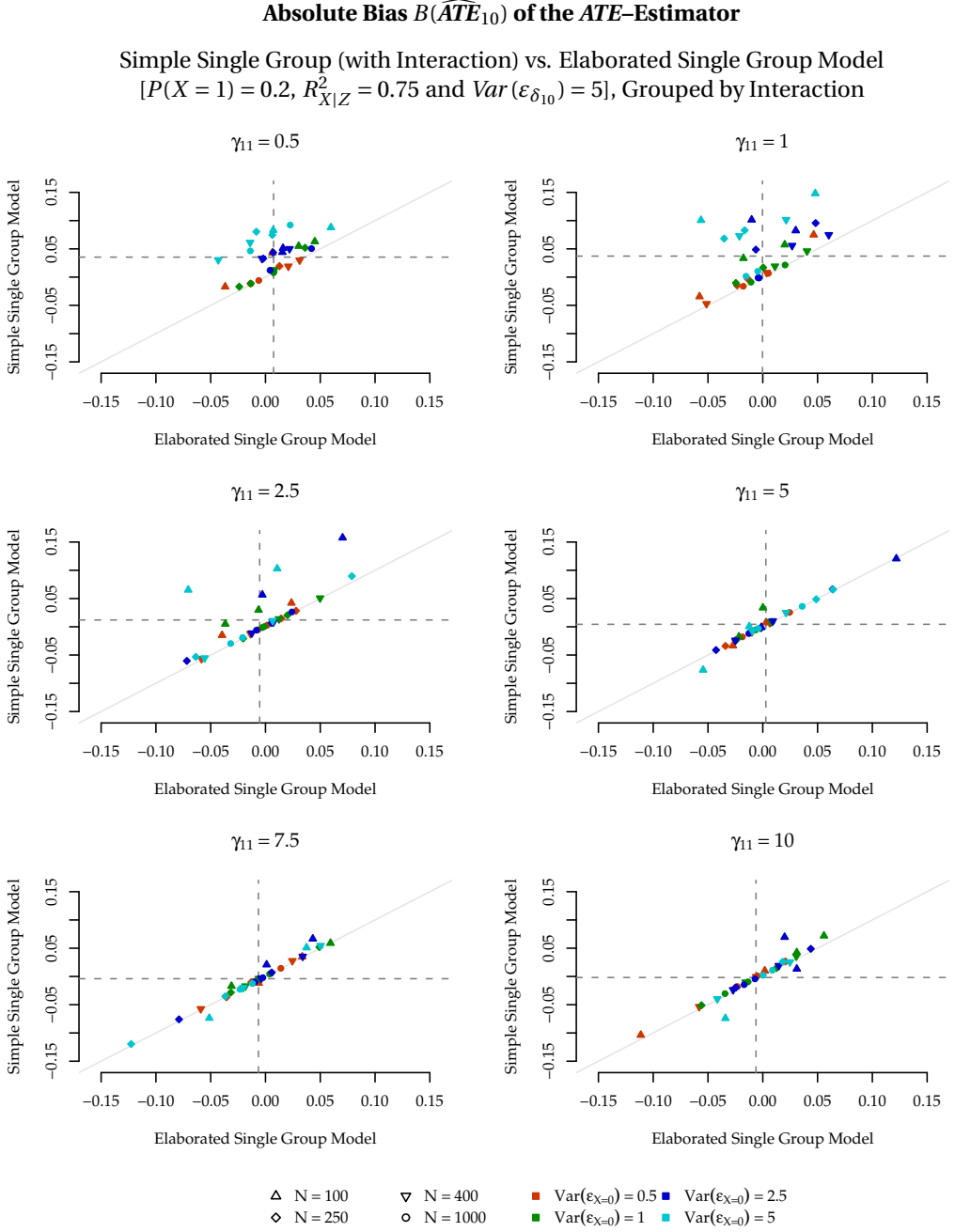


Figure 4.37: Absolute bias of the *ATE*–estimator: Scatter plots for a comparison of the simple single group model (with interaction) and the elaborated single group model, grouped by γ_{11} [$R^2_{X|Z} = 0.75$, $P(X = 1) = 0.8$ and $Var(\varepsilon_{\delta_{10}}) = 5$]

interaction parameter γ_{11}). Note that the *ATE*–estimator's absolute bias for $P(X = 1) = 0.8$ is largest for conditions with the smallest level of the interaction parameter used for data generation ($\gamma_{11} = 0.5$). The estimator's bias is almost zero for medium interaction effects ($\gamma_{11} = 2.5$), and the observed biases are again slightly higher for interaction parameters greater than five ($\gamma_{11} \geq 5$).

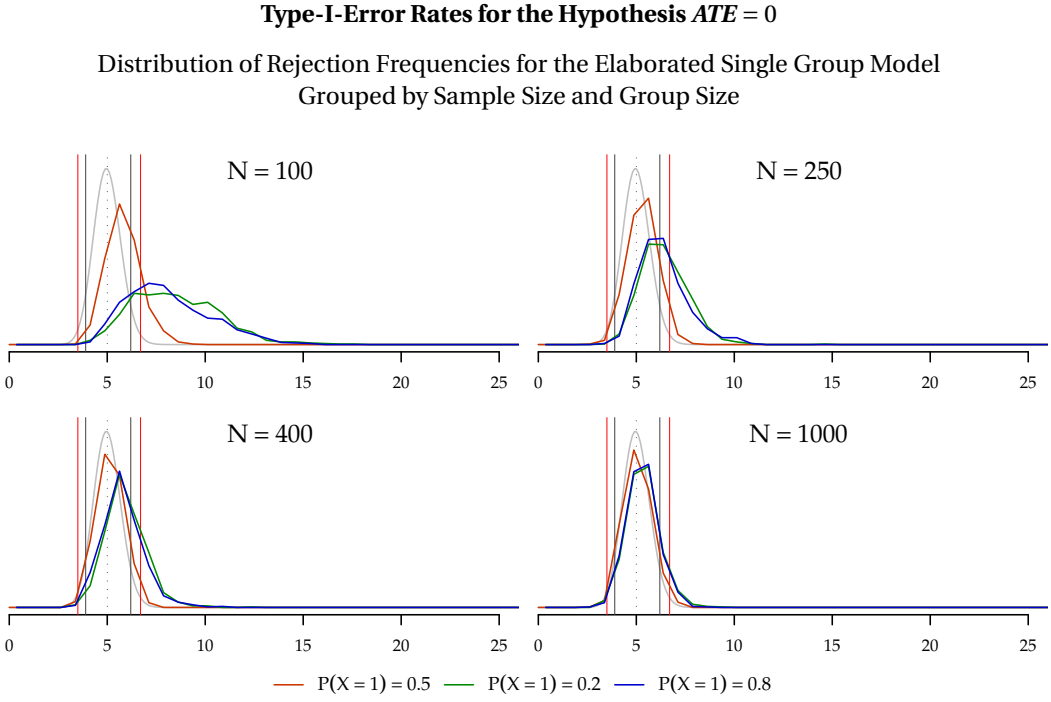


Figure 4.38: Type-I-error rate: Distribution of rejection frequencies for the elaborated single group model, grouped by sample size N and group size $P(X = 1)$

Type-I-Error Rate We now turn to the empirical type-I-error rates for the elaborated single group model compared to the simple single group model, i. e., we will compare the observed distributions of the rejection frequencies for the test of the hypothesis $ATE_{10} = 0$ for both potential implementations of generalized analysis of covariance. Acceptable rejection frequencies for equal group sizes are obtained from the elaborated single group model for moderate sample sizes ($N \geq 400$), while for unequal group sizes the method yields h_{RF}'s within the confidence intervals only for large sample sizes ($N = 1000$). For sample sizes smaller than $N = 400$ the small sample behavior of the Wald–test based on the elaborated single group model results in a systematic inflation of the empirical type-I-error rates. An inspection of Figure 4.38 reveals that the distributions of h_{RF}, in contrast to the simple single group model,[53] are almost identical for unequal group sizes with treatment probabilities of $P(X = 1) = 0.2$ and $P(X = 1) = 0.8$. Hence, the results in general confirm the theoretically motivated extension of the simple single group model with respect to the second order missspecification of the implied variance structure.

[53] For the simple single group model, the observed distributions of h_{RF} for all cells of simulation study I, grouped by sample size and plotted as separate lines for the three group size conditions are presented as additional Figure 64 on page 75 of the digital appendix. The expected distribution (thin gray line, see also section 4.4.3 for the concrete meaning) is well approximated only for some selected conditions, namely for equal group sizes and sample sizes larger than $N = 250$ (see red lines in the lower two of the four line charts in the additional Figure 64).

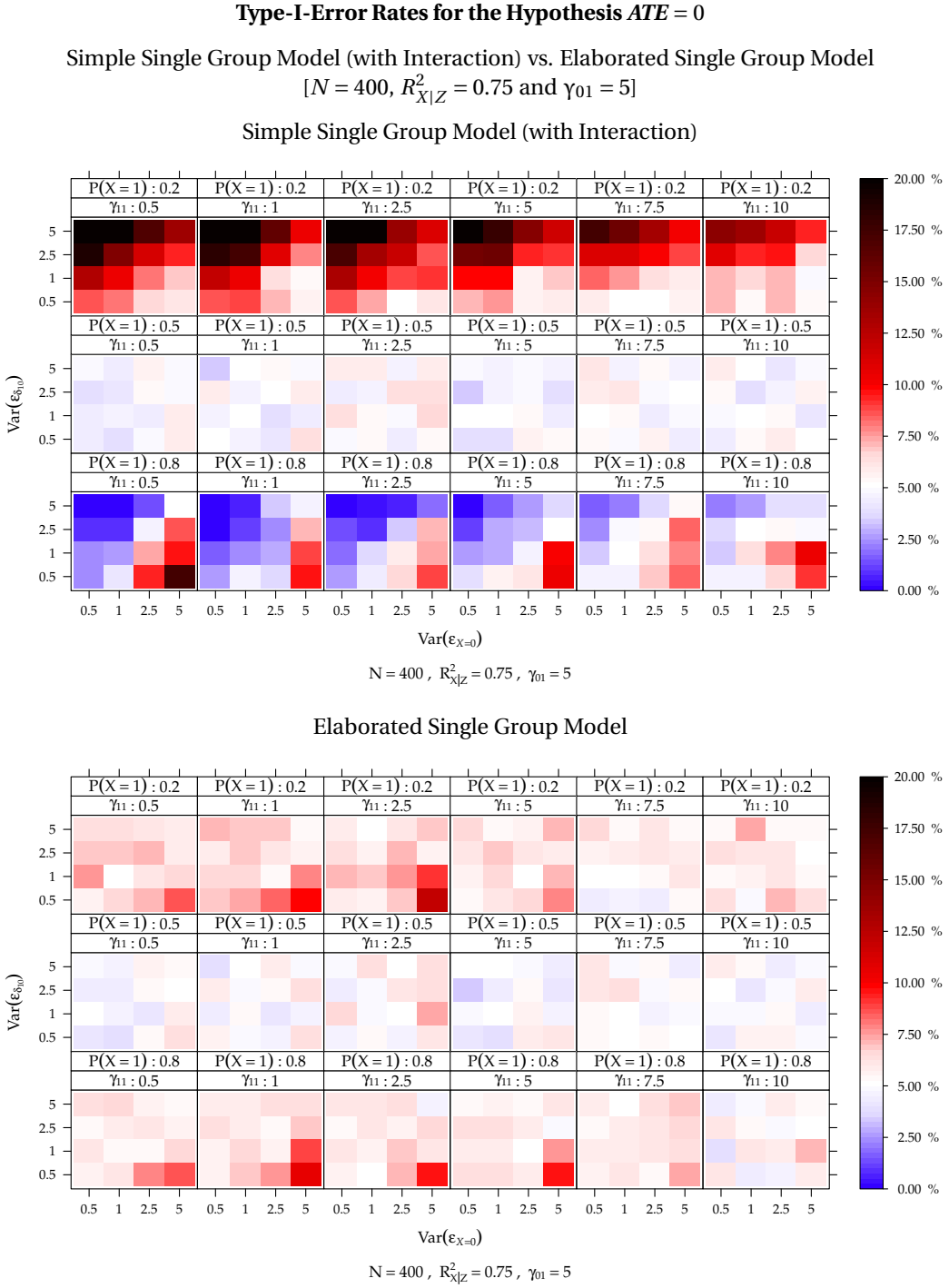


Figure 4.39: Type-I-error rate: Level plots for a comparison of the simple single group model (with interaction) and the elaborated single group model [$N = 400$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

A critical inspection of the lower part of Figure 4.39 gives reason to take a closer look at the behavior of the elaborated single group model for conditions with large values of the residual variance $Var(\varepsilon_{X=0})$ and very small amounts of the residual variance $Var(\varepsilon_{\delta_{10}})$. Unfortunately, the elaborated single group model is

sensitive to a second order misspecification of the implied variance structure as well as the simple single group model. We observe systematically inflated empirical type-I-error rates for medium and small sample sizes under conditions of the simulation study I where the residual variances are almost homogeneous.[54] Note that this pattern of inflated empirical type-I-error rates is distinct from the pattern of the absolute biases of the *ATE*–estimator.[55]

Furthermore, we observe an overall pattern of inflated empirical type-I-error rates for small values of the interaction parameter γ_{11}. The elaborated single group model needs larger sample sizes to achieve an acceptable empirical type-I-error rate for simulated conditions with small interaction effects (compared to conditions with large interaction effects). This finding is most obvious for unequal group sizes (see, for example, the first and the third row in Figure 4.39).[56] For large sample sizes the rejection frequencies are acceptable for equal group sizes as well as for unequal group sizes. For small sample sizes ($N = 100$) the empirical type-I-error rates are slightly inflated for datasets generated with small interaction parameters ($\gamma_{11} < 5$).

Convergence Rate The finding that the elaborated single group model is sensitive to a misspecified implied variance structure can also be verified by an inspection of the convergence rates. A comparison of the convergence rates for the simple single group model and the elaborated single group model in Figure 4.40 reveals the following interesting phenomenon: Although the elaborated single group model converges much better for unequal group sizes [when the treatment groups are larger than the control groups, $P(X = 1) = 0.8$] than the simple single group model, the convergence rates of the elaborated single group model break down systematically for conditions with a large residual variance [$Var(\varepsilon_{X=0})$, in particular for datasets generated with a small residual variance of the individual total effect, $Var(\varepsilon_{\delta_{10}})$]. This finding is in line with the effect of the misspecified implied variance structures on the observed rejection frequencies. Obviously, conditions under which the assumption of homogeneity of between-group residual variances almost holds are most critical for the elaborated single group model.[57]

Bias of the Standard Error of the *ATE*–Estimator An inspection of the relative biases of the standard error of the average total effect estimator confirms the findings already presented for the empirical type-I-error

[54] In addition to Figure 4.39 we present the empirical type-I-error rates as level plots for $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$ in the digital appendix as the additional Figure 65 on page 76 for $N = 1000$, as the additional Figure 66 on page 77 for $N = 250$, and particularly as the additional Figure 67 on page 78 for $N = 100$.

[55] This follows from the fact that we can observe the systematic pattern for the empirical type-I-error rates in Figure 4.39, but not for the corresponding plot of the $B(\widehat{ATE}_{10})$, presented as the additional Figure 68 on page 79 of the digital appendix.

[56] See also the additional figures mentioned in footnote 54.

[57] The additional Figure 69 on page 80 of the digital appendix presents the corresponding density plots of the observed convergence rates conditional on the residual variance of the individual total effect $Var(\varepsilon_{\delta_{10}})$. Within the different plots the color of the symbols refers to the factor $Var(\varepsilon_{X=0})$. The convergence rates of the elaborated single group model are inconspicuous for large values of the residual variances of the individual total effect and worst for conditions where this variance is small (and the residual variance in the control group is high).

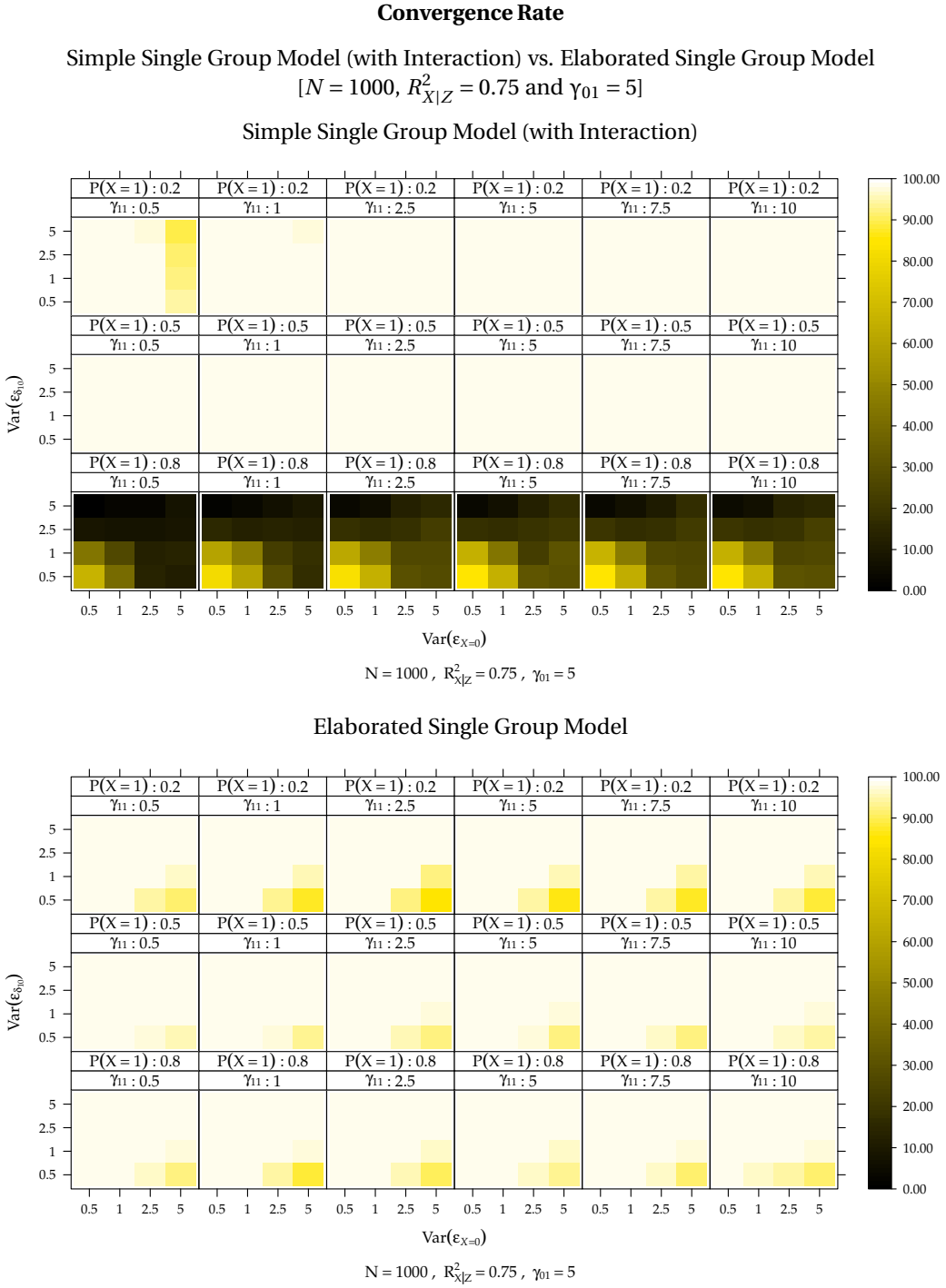


Figure 4.40: Convergence rate: Level plots for a comparison of the simple single group model (with interaction) and the elaborated single group model [$N = 1000$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

rates and for the convergence rates: The $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ obtained from the elaborated single group model is almost zero for large sample sizes in most conditions of simulation study I. Sound biases are observed under some conditions where the individual total effect is strongly determined by the covariate [that means

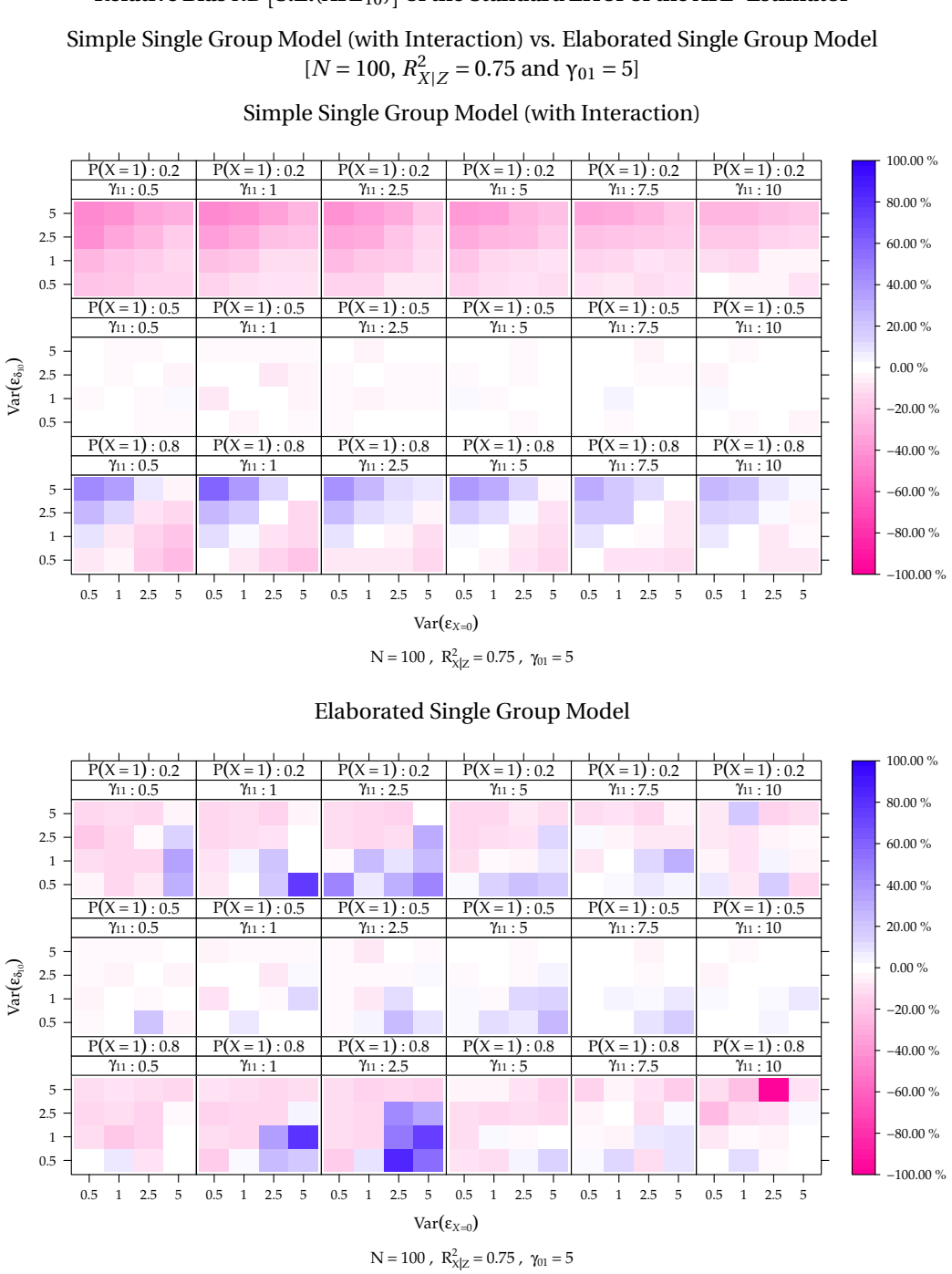


Figure 4.41: Relative bias of the standard error of the *ATE*–estimator: Level plot for a comparison of the simple single group model (with interaction) and the elaborated single group model [$N = 100$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

where $Var(\varepsilon_{\delta_{10}})$ is small compared to $Var(\varepsilon_{X=0})$]. As seen in the lower part of Figure 4.41, the observed pattern of the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for unequal group sizes mirrors to some extent the biases of the *ATE*–estimator's

standard error obtained from the simple single group model.[58] A comparison of the observed pattern for $P(X = 1) = 0.8$ and $\gamma_{11} = 2.5$ (see the third column in the third row of Figure 4.41) suggests that standard errors obtained from the simple single group model are positively biased for heterogenous residual variances, whereas for the elaborated single group model negatively biased standard errors can be observed in the same generated data sets with homogenous residual variances.[59]

Summary The elaborated single group model, developed as an extension of the simple single group model performed adequately, at least for large sample sizes and when the individual total effect was not completely determined by the covariate (i. e., for conditions with heterogeneity of residual variance). Due to the additional random slope residual, the implied variance structure of this extended single group model matched the structure used for generating the data under heterogeneity of residual variances. The observed performance was not satisfying for small sample sizes, especially for unequal group sizes. Under these conditions we observed again consequences of the misspecified implied variance structure. The biased standard errors of the average total effect estimator and the observed convergence problems occurred antipodal to the failures of the simple single group model (with interaction), if the residual variances of the covariate-treatment regression were generated homogenous. The standard error of the *ATE*–estimator was underestimated by the simple single group model for some of the considered conditions, while the standard error was overestimated by the elaborated single group model for exactly the same data. Therefore, we conclude that the single group approach for interaction modeling within the framework of structural equation modeling seems to be more fragile than the multi-group approach, meaning that the parameter estimates and the estimated asymptotic variances and covariances are generally sensitive to (second order) model misspecifications. This is an important finding, as for interactions of continuous covariates $Z = (Z_1, ..., Z_K)$ with a dichotomous treatment variable X the multi-group modeling approach becomes an even more interesting choice.

[58] See also the additional Figure 70 on page 81 for the results for $N = 250$, the additional Figure 71 on page 82 for the results for $N = 400$ and the additional Figure 72 on page 83 for the results for $N = 1000$. For larger sample sizes the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ of the *ATE*–estimator obtained from the elaborated single group model disappear completely.

[59] Detailed views on the observed relative standard error biases conditional on the two residual variances $Var(\varepsilon_{X=0})$ and $Var(\varepsilon_{\delta_{10}})$ are presented in the additional Figure 76 on page 87, Figure 77 on page 88 and Figure 78 on page 89 of the digital appendix. In each of the three figures 16 scatter plots of the relative bias of the standard error for the *ATE*–estimator obtained from the elaborated single group model (y–axis) and from the simple single group model (x–axis) are printed for the sample size of $N = 1000$. The scatter plots are grouped by the value of the factor $Var(\varepsilon_{X=0})$ [vertical] and the different amounts of $Var(\varepsilon_{\delta_{10}})$ [horizontal] used for generating the data, in order to visualize the impact of these two residual variances. The standard errors obtained from the elaborated single group model are biased for small values of $Var(\varepsilon_{\delta_{10}})$ and large values of $Var(\varepsilon_{X=0})$ for equal group sizes (see first column and forth row in Figure 76). In general the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ of the elaborated single group models is remarkably smaller than the bias obtained from the simple single group model (i. e., the plotted data points drift off the diagonal). A similar drift can be observed for equal group sizes (Figure 76) and for unequal group sizes (Figure 77, and Figure 78), but in addition the effect of the misspecified implied variance for the simple single group model dominates the picture: For large values of $Var(\varepsilon_{\delta_{10}})$ and small values of $Var(\varepsilon_{X=0})$ [see the forth column in the first row of both figures], the standard error for the simple single group model is heavily biased. Furthermore, note that the observed bias of the standard error of the *ATE*–estimator for generated datasets are not connected to the sample size when the amount of residual variance heterogeneity is small. This finding can be confirmed by comparing the presented relative standard error biases in Figure 76, Figure 77, and Figure 78 to the corresponding plots for generated datasets with $N = 100$, provided in the digital appendix (see Figure 73, Figure 74 and Figure 75 on page 84 et seqq.).

4.7.2 Elaborated Multi-Group Model

The next model we are analyzing in more detail is the *elaborated multi-group model*. For the investigation of different statistical procedures for generalized analysis of covariance within the framework of structural equation models this approach is in a prominent position. On the one hand it is expected that the elaborated multi-group model – in contrast to the simple multi group model — results in unbiased standard errors for the *ATE*–estimator. Moreover, as already reported, all multi-group yielded an unbiased estimator of the average total effect. On the other hand the results of the simulation study for this implementation of generalized analysis of covariance will be informative also with respect to the assumption of uncorrelated parameter estimates (see the research question in subsection 3.4.5).

Type-I-Error Rate The implementation of generalized analysis of covariance as structural equation model with nonlinear constraint based on the `knownclass`–option of `Mplus` yields, as expected, acceptable empirical type-I-error rates for large sample sizes. The observed distribution of rejection frequencies for tests of the hypothesis $ATE = 0$ over all conditions of simulation study I for $N = 1000$ is almost equal to the expected distribution for a nominal type-I-error rate of 5 % (see Figure 4.42).[60]

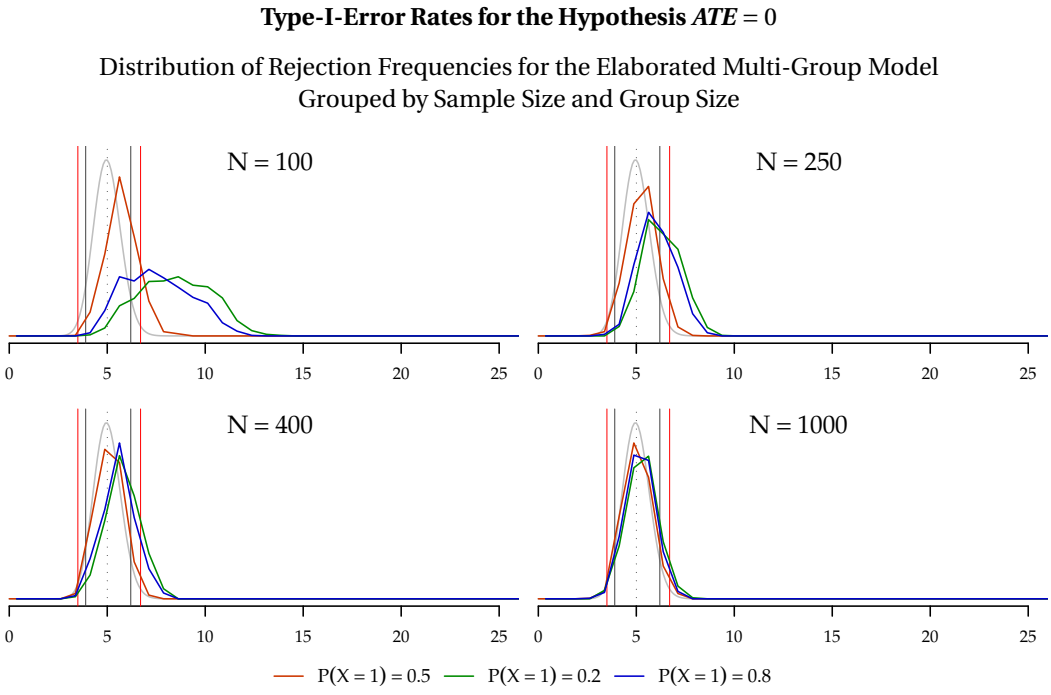


Figure 4.42: Type-I-error rate: Distribution of rejection frequencies for the elaborated multi-group model, grouped by sample size N and group size $P(X = 1)$

[60] See also the upper of the level plot presented as additional Figure 79 on page 90 of the digital appendix.

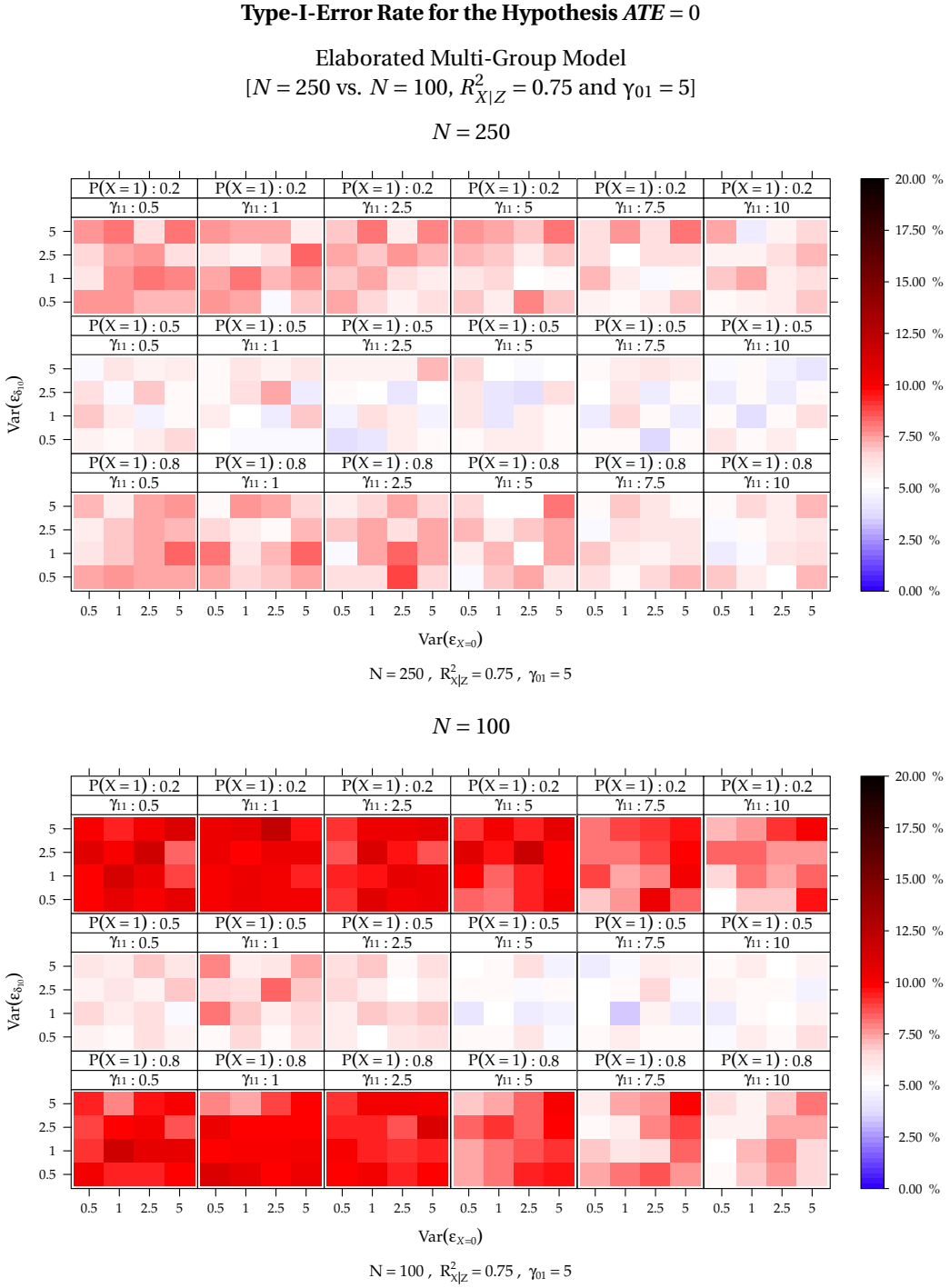


Figure 4.43: Type-I-error rate: Level plot for the elaborated multi-group model [$N = 250$ and $N = 100$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

The rejection frequencies for the Wald–test of the hypothesis $ATE = 0$ observed for the elaborated multi-group model are much more reasonable for unequal group sizes, and medium and large sample sizes, and in particular as compared with the empirical type-I-error rates observed for the simple multi-group

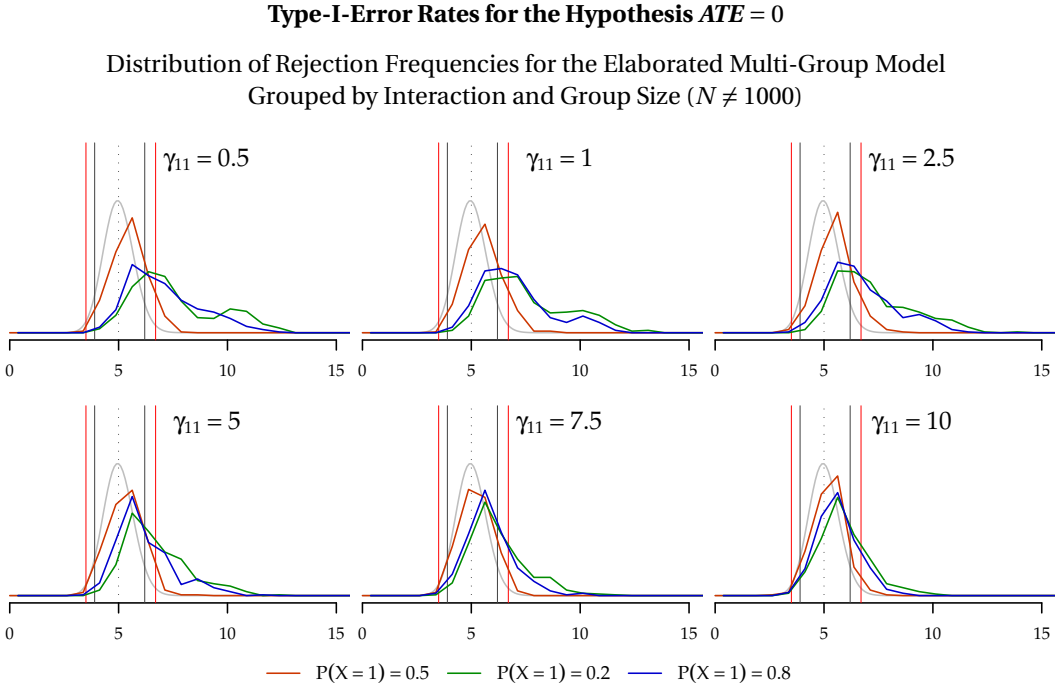


Figure 4.44: Type-I-error rate: Distribution of rejection frequencies for the elaborated multi-group model, grouped by interaction γ_{11} and group size $P(X = 1)$ [$N \neq 1000$]

model.[61] Nevertheless, the averaged distributions of the rejection frequencies in Figure 4.42 are noticeably expanded for datasets with unequal group sizes and are clearly not around the nominal level of 5 % for all conditions of the simulation study I with small sample sizes ($N = 100$).

Furthermore, it is interesting to note that, for datasets generated with small interaction parameters, the elaborated multi-group model results in inflated empirical type-I-error rates for most considered sample sizes, which decrease to the nominal level as the interaction parameter γ_{11} increases. This pattern is most obvious for medium sample sizes and treatment probabilities different from $P(X = 1) = 0.5$ (see the upper part of Figure 4.43), as well as for equal group sizes and small sample sizes (see the lower part of Figure 4.43). To visualize this behavior of the elaborated multi-group model more explicitly, the distributions of the observed rejection frequencies for all conditions of simulation study I with $N \neq 1000$ are plotted in Figure 4.44 conditional on the levels of the interaction parameter γ_{11}: Whereas the observed distribution follows approximately the expected form for equal group sizes (around the nominal level and within the confidence intervals, see the red lines in Figure 4.44), the observed distribution of the empirical type-I-error rates over all conditions of simulation study I with $N \neq 1000$ for unequal group sizes are shifted for small

[61] See the additional Figure 80 on page 91 of the digital appendix.

values of γ_{11} (see the distributions plotted in green and blue).[62] The nominal type-I-error rate is achieved for all studied levels of the interaction parameter γ_{11} for large sample sizes.[63]

Bias of the Standard Error of the *ATE*–Estimator The standard error of the average total effect estimator is negatively biased with −2.5 % (average over all conditions of simulation study I). A bias smaller than one percent is observed for the elaborated multi-group model (−0.77 %) for conditions with a treatment probability equal to $P(X = 1) = 0.5$, whereas the simple multi-group model (based on the estimated group size, sample) is biased larger than Florys' selected 5 % cut-off value for those conditions (−6.71 %). Table 4.8 summarizes the observed relative biases of the *ATE*–estimator for the simple multi-group model (sample) and the elaborated multi-group model, separately for the four different sample sizes and the three different group sizes considered in simulation study I. As presented in the table, both structural equation models with nonlinear constraints give biased standard errors for conditions with $N = 100$. The relative advantage of the elaborated multi-group model in terms of $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ is obvious for equal treatment probabilities as well as for medium and large sample sizes ($N \geq 250$).

Table 4.8: Relative bias of the standard error of the *ATE*–estimator: Simple multi-group model (sample) and elaborated multi-group model, grouped by sample size N and group size $P(X = 1)$

		Elaborated Multi-Group Model			Simple Multi-Group Model (Sample)		
N	$P(X = 1)$	0.5	0.2	0.8	0.5	0.2	0.8
100		−1.76 %	**−8.79** %	**−7.03** %	**−7.6** %	**−8.81** %	**−8.27** %
250		−0.8 %	−3.39 %	−2.67 %	**−6.69** %	−4.85 %	**−5.45** %
400		−0.35 %	−2.01 %	−1.6 %	**−6.35** %	−4.04 %	−4.56 %
1000		−0.15 %	−0.84 %	−0.6 %	**−6.2** %	−3.13 %	−3.93 %

Note: The hypothesis of no average total effect based on the simple multi-group model (sample) is tested using the estimated group sizes in the nonlinear constraint.

The relative biases of the *ATE*–estimator's standard error for all conditions with $N = 1000$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$ are summarized as level plot in Figure 4.45 (the level plot in the upper part of the figure refers to the elaborated multi-group model, and the plot in the lower part refers to the simple multi-group model based on the sample estimate of the group size). For the elaborated multi-group model and large sample sizes a comparable pattern of the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ can be observed as previously described for the observed rejection frequencies under medium sample sizes (see last paragraph).[64]

[62] Similar figures are provided for the approximated multi-group model, see the additional Figure 81 on page 92 of the digital appendix and for the elaborated single group model, see the additional Figure 82 on page 93 of the digital appendix. Note that — because the interplay between the interaction parameter γ_{11} and the empirical type-I-error rate is observable only for small sample sizes — we excluded all conditions with $N = 1000$ for generating these plots.

[63] The small sample behavior of the elaborated multi-group model, along with all other implementations of generalized analysis of covariance, is studied in detail based on part II of the Monte Carlo simulation, presented in section 4.8.

[64] Note that we in fact studied two different standard errors for the *ATE*–estimator based on the elaborated multi-group model: The δ–theorem applied to the estimated variance-covariance matrix of parameter estimates as obtained from the elaborated multi-group model, and the δ–theorem applied to the same matrix incorporating the assumption of uncorrelated parameter estimates as used for the approximated multi-group model. Both strategies yield indistinguishable results for the average of the standard errors for all cells of the simulation study. Hence, we do not present the results based on the two slightly different standard errors for the *ATE*–estimator. In the digital supplement the different methods are labeled as `A` for the approach presented here and `C` for the method where the additional asymptotic covariances between the estimate of the group size and all other model parameters are set to zero.

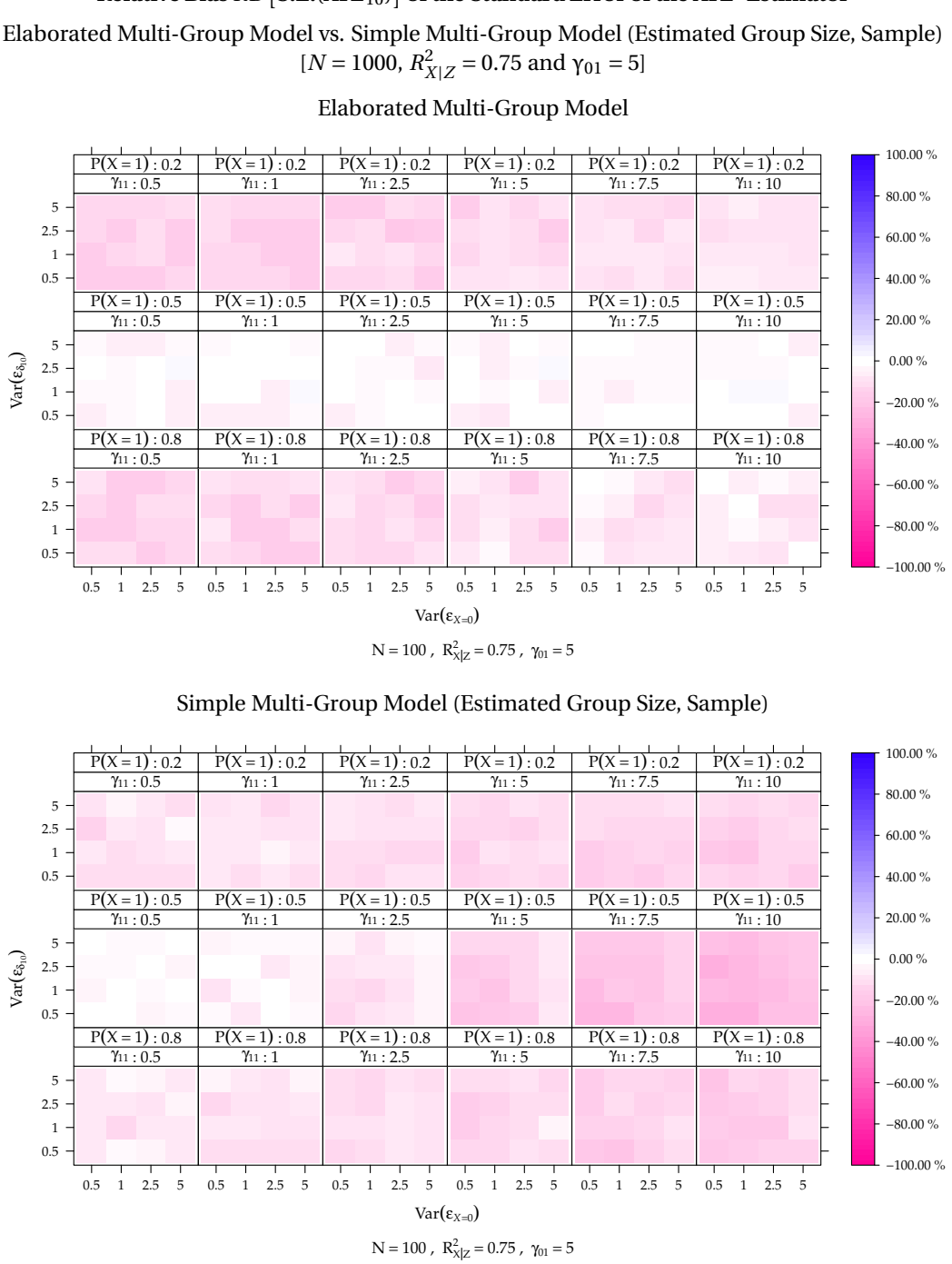


Figure 4.45: Relative bias of the standard error of the *ATE*–estimator: Level plots for the elaborated single group model and the simple multi-group model (sample) [$N = 1000$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

The overall performance of the elaborated multi-group model in terms of $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$, as compared to the observed biases of the standard error of the *ATE*–estimator for the simple multi-group model based on the estimated group size, appears in Table 4.9, conditional on the interaction parameter γ_{11} and

Table 4.9: Relative bias of the standard error of the *ATE*–estimator: Elaborated multi-group model and the simple multi-group model (sample), grouped by interaction γ_{11}

			Elaborated Multi-Group Model			Simple Multi-Group Model (Sample)		
N	γ_{11}	$P(X=1)$	0.5	0.2	0.8	0.5	0.2	0.8
100	0.5		−2.49 %	**−11.62** %	**−10.21** %	−2.1 %	**−8.06** %	**−6.79** %
250			−0.97 %	−4.83 %	−3.75 %	−1.16 %	−2.92 %	−2.77 %
400			−0.68 %	−2.41 %	−2.01 %	−0.42 %	−1.72 %	−1.43 %
1000			−0.51 %	−1.3 %	−0.87 %	−0.5 %	−0.5 %	−0.66 %
100	1		−2.39 %	**−11.58** %	**−9.8** %	−2.78 %	**−8.1** %	**−6.79** %
250			−0.86 %	−4.34 %	−3.66 %	−1.67 %	−2.79 %	−2.44 %
400			−0.3 %	−3.16 %	−2.16 %	−1.29 %	−2.24 %	−1.85 %
1000			−0.13 %	−1.27 %	−0.89 %	−0.71 %	−0.8 %	−0.69 %
100	2.5		−2.21 %	**−10.33** %	**−8.46** %	**−5.43** %	**−8.28** %	**−7.39** %
250			−1.34 %	−4.35 %	−3.26 %	−4.03 %	−3.64 %	−4.03 %
400			−0.62 %	−2.21 %	−2.02 %	−3.94 %	−2.78 %	−2.66 %
1000			−0.32 %	−0.81 %	−1.14 %	−3.86 %	−1.58 %	−2.27 %
100	5		−1.58 %	**−8.47** %	**−5.94** %	**−9.5** %	**−8.63** %	**−8.65** %
250			−0.48 %	−3.17 %	−2.19 %	**−8.69** %	**−5.23** %	**−6.06** %
400			−0.24 %	−1.98 %	−1.36 %	**−8.2** %	−4.45 %	**−5.34** %
1000			0.04 %	−0.92 %	−0.36 %	**−8.2** %	−3.73 %	−4.86 %
100	7.5		−1.28 %	**−5.99** %	−4.47 %	**−11.92** %	**−9.64** %	**−9.6** %
250			−0.53 %	−1.9 %	−1.62 %	**−11.4** %	**−6.61** %	**−7.96** %
400			−0.24 %	−1.24 %	−1.25 %	**−11.2** %	**−5.64** %	**−7.33** %
1000			0.15 %	−0.17 %	−0.37 %	**−11.41** %	**−5.36** %	**−6.75** %
100	10		−0.64 %	−4.73 %	−3.29 %	**−13.88** %	**−10.13** %	**−10.4** %
250			−0.66 %	−1.73 %	−1.54 %	**−13.19** %	**−7.91** %	**−9.45** %
400			−0.01 %	−1.07 %	−0.8 %	**−13.07** %	**−7.41** %	**−8.74** %
1000			−0.13 %	−0.54 %	0.01 %	**−12.53** %	**−6.78** %	**−8.37** %

Note: The hypothesis of no average total effect based on the simple multi-group model (sample) is tested using the estimated group sizes in the nonlinear constraint.

separately for the different levels of the sample size N and for the different treatment probabilities $P(X=1)$. The observed standard errors are biased for the elaborated multi-group model as well as for the simple multi-group model for the smallest sample size considered in the first part of the simulation study ($N=100$) and for small and medium interactions effects ($\gamma_{11}<7.5$) and unequal group sizes. Nevertheless, the structure of these biases differs between both models, as the observed underestimation of the standard error of the *ATE*–estimator obtained from the elaborated multi-group model decreases with the amount of interaction γ_{11} used for generating the data.[65] Comparisons of the relative biases obtained for the simple multi-group model (sample) to the biases obtained for the elaborated multi-group model (see Figure 4.45 and Table 4.8 and Table 4.9) confirm again the argumentation that the mistreatment of the stochasticity of X causes the bias of the standard error of the average total effect estimator for the simple multi-group model (sample) for large interaction effects (see subsection 4.6.1).

[65]Actually, the small sample performance of the elaborated single group model is superior to the to the performance observed for the simple multi-group model (sample) for large interaction effects. For $N=400$, $N=250$ and $N=100$ this is plotted in the digital appendix in Figure 83, Figure 84 and Figure 85 on page 94 and the following.

Asymptotic Covariances of Parameter Estimates The central assumption of uncorrelated parameter estimates underlying the *approximated multi-group model* was introduced and described in detail in subsection 3.3.3.3. For the computation of the standard error of the *ATE*–estimator for the approximated multi-group model, prior to the application of the δ–theorem, an augmentation of the variance-covariance matrix of parameter estimates is performed. This augmentation is based on the assumption that the estimates of all model parameters of the structural equation model are uncorrelated with the estimate(s) of the group size. The assumption of zero asymptotic covariances between the corresponding parameter estimates is not necessary for the elaborated multi-group model. Furthermore, we obtain estimates of the critical asymptotic covariances when estimating the elaborated multi-group model because $P(X = 1)$ [or more precisely $E(C)$] is estimated as an additional model parameter. As described in subsection 3.3.3.2 (see page 92), a transformation of these estimated covariances between $E(C)$ and all other model parameters into the focused covariances between $P(X = 1)$ and the estimated parameters of the structural equation model is possible with the help of the multivariate δ-method.

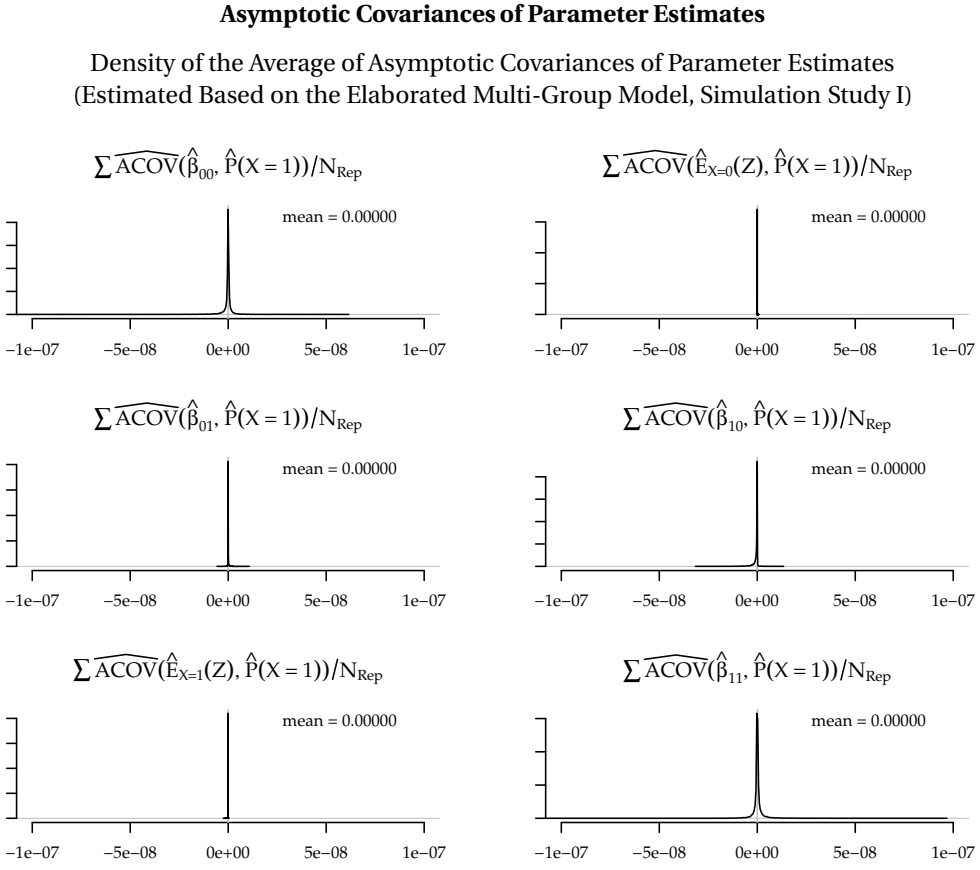


Figure 4.46: Densities of the average of asymptotic covariances of parameter estimates, estimated based on the elaborated multi-group model (simulation study I)

Figure 4.46 presents density plots for each of these disputable asymptotic covariances (averaged within each cell of simulation study I), based on the estimates obtained from the elaborated multi-group model. Obviously, those asymptotic covariances are zero across all conditions of simulation study I (no average total effect).[66]

Furthermore, we also analyzed the empirical covariances between parameter estimates. As described for the standard errors of the *ATE*–estimator (see paragraph 4.4.2), the empirical covariances of parameter estimates are expected to be equal to the asymptotic covariances between parameter estimates as used for the δ-method. In order to flesh out the assumption underlying the approximated multi-group model, these empirical covariances are plotted in Figure 4.47.

Empirical Covariances of Parameter Estimates

Density of the Empirical Covariances of Parameter Estimates
(Estimated Based on the Elaborated Multi-Group Model, Simulation Study I)

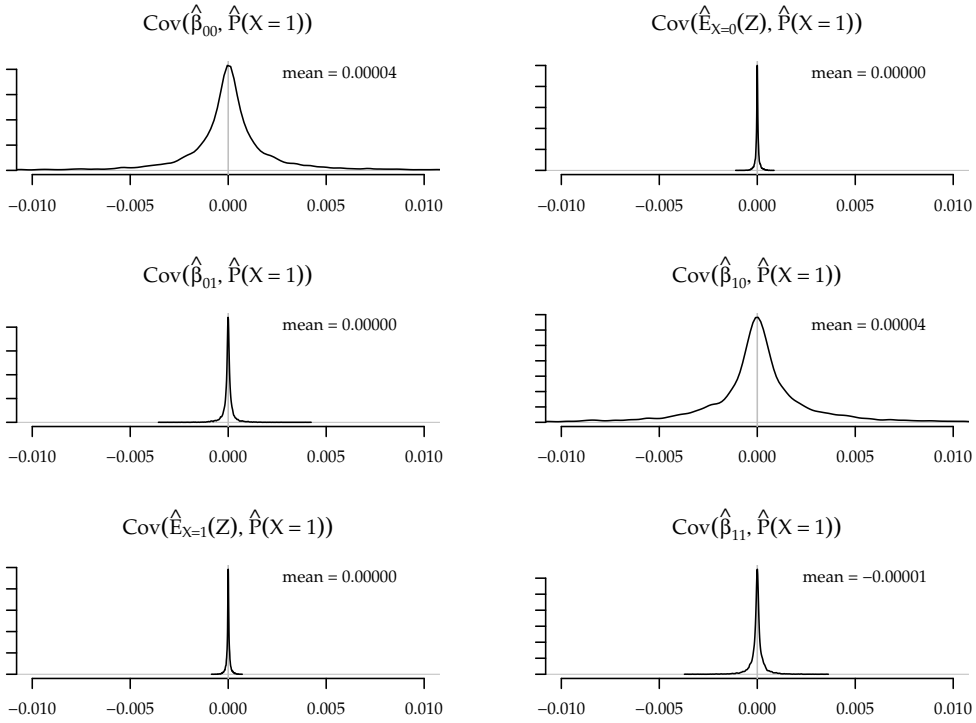


Figure 4.47: Density of the empirical covariances of parameter estimates, estimated based on the elaborated multi-group model (simulation study I)

Both measures support the assumption that the covariances between estimated parameters of the multi-group structural equation model and the estimated group size are zero. Since the simple multi-group model results in biased standard errors of the *ATE*–estimator and inflated empirical type-I-error rates for

[66] The same result is obtained for part II of the simulation study (for small samples and average total effects with effect sizes varying from $d = 0$ to $d = 0.8$, see the additional Figure 86 on page 97 of the digital appendix)

the test of the hypothesis $ATE = 0$ we conclude that the stochasticity of the group size, i. e., the asymptotic variance of the estimated treatment probability must be considered.

Summary The developed elaborated multi-group modeling approach (i. e., the multi-group model with group size as estimated model parameter based on the `knownclass`–option of `Mplus`) performed generally well. Only for small sample sizes ($N = 100$) and unequal group sizes, did the simple multi-group model (sample) produce a distribution of rejection frequencies closer to the expected distribution than the elaborated multi-group model. One of the most important findings is that the elaborated multi-group model performs equally well regardless of the heterogeneity of residual variance [i. e., for all levels of $Var(\varepsilon_{\delta_{10}})$]. This points our attention to further research on the technical differences in the model estimation of the different structural equation models, in particular on the differences between the estimation of the single group models versus the estimation of the multi-group models. Finally, the uncorrelated parameter estimates can be interpreted as empirical evidence for the appropriateness of the augmentation-based approach as described in subsection 3.3.3.3. The performance of the corresponding approximated multi-group model is presented in the next subsection.

4.7.3 Approximated Multi-Group Model

The fifth structural equation model we studied in the Monte Carlo simulation is the *approximated multi-group model* based on the augmented variance-covariance matrix of parameter estimates. Nagengast (2006) reported acceptable empirical type-I-error rates and unbiased standard errors of the *ATE*–estimator for the augmentation-based approach under simulated conditions with homogeneity of between-group residual variances. Based on the results presented in this section we will try to generalize the appropriateness of this approach to conditions with residual variance heterogeneity. Therefore, we compare the empirical type-I-error rates and especially the standard errors of the average total effect estimator to the results obtained from the elaborated multi-group model.

Table 4.10: Type-I-error rate: Approximated multi-group model and elaborated multi-group model, grouped by sample size N and group size $P(X = 1)$

		Approximated Multi-Group Model			Elaborated Multi-Group Model		
N	$P(X = 1)$	0.5	0.2	0.8	0.5	0.2	0.8
100		5.54 %	6.94 %	6.46 %	5.69 %	8.39 %	7.59 %
250		5.23 %	5.65 %	5.59 %	5.32 %	6.32 %	6.00 %
400		5.12 %	5.41 %	5.30 %	5.15 %	5.83 %	5.61 %
1000		5.07 %	5.13 %	5.10 %	5.08 %	5.35 %	5.27 %

Type-I-Error Rate The central results concerning the key assumption underlying the augmentation approach used for the *approximated multi-group model* were presented in the previous subsection. We found no sound reason for assuming asymptotic covariances between parameter estimates of the group size

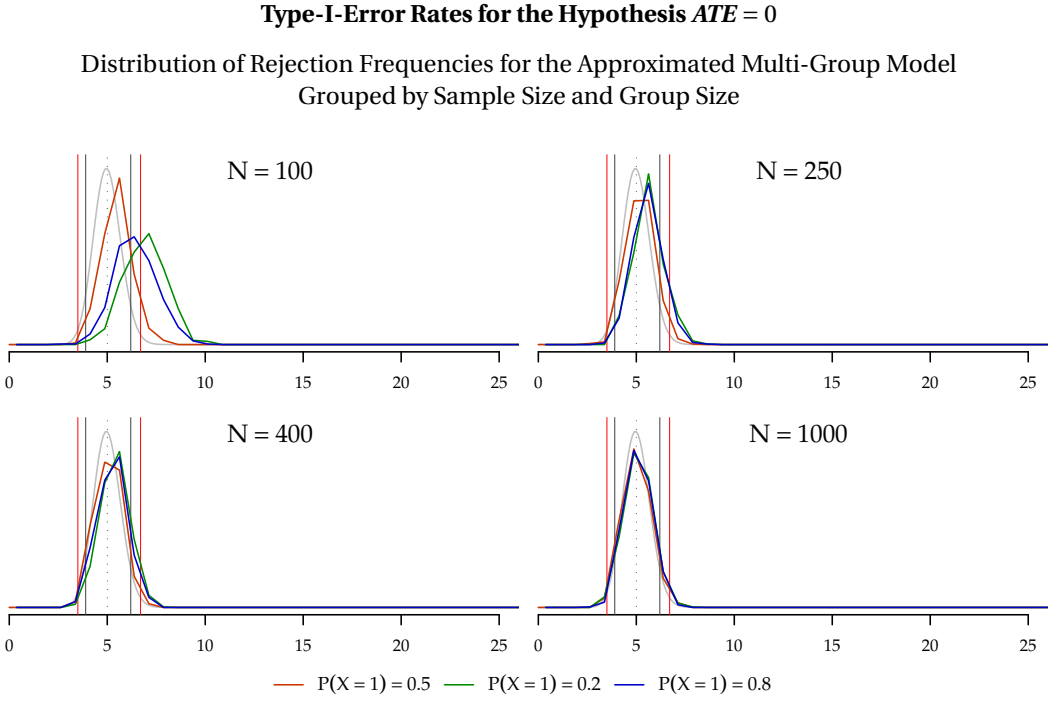


Figure 4.48: Type-I-error rate: Distribution of rejection frequencies for the approximated multi-group model, grouped by sample size N and group size $P(X = 1)$

$P(X = 1)$ and all remaining model parameters of the multi-group covariate-treatment regression. Therefore, it is expected that the approximated multi-group model displays rejection frequencies for the hypothesis of no average total effect at the nominal level, i. e., empirical type-I-error rates almost identical to the results observed for the elaborated multi-group model.

Table 4.11: Type-I-error rate: Approximated multi-group model and elaborated multi-group model, grouped by dependency between X and Z $R^2_{X|Z}$ and group size $P(X = 1)$

		Approximated Multi-Group Model			Elaborated Multi-Group Model		
$R^2_{X\|Z}$	$P(X = 1)$	0.5	0.2	0.8	0.5	0.2	0.8
0.75		−0.66 %	−3.03 %	−2.7 %	−1.04 %	**−5** %	−4.3 %
0.5		−0.62 %	−2.8 %	−2.07 %	−0.86 %	−4.45 %	−3.48 %
0.25		−0.54 %	−2.43 %	−1.83 %	−0.75 %	−3.35 %	−2.63 %
0.1		−0.63 %	−1.68 %	−1.18 %	−0.41 %	−2.22 %	−1.49 %

The empirical type-I-error rates for the approximated multi-group model for all conditions of part I of the simulation study confirm our expectation: 5.24 % of the tests for generated datasets with equal group sizes and no true average total effect were significant [for $P(X = 1) = 0.2$ on average 5.78 % and for $P(X = 1) = 0.8$ overall 5.61 % significant tests were observed].[67] The empirical distributions of the rejec-

[67] We observed for the elaborated multi-group model 5.31 % significant tests on average for equal group sizes. For treatment probabilities of $P(X = 1) = 0.2$, on average 6.47 % of the tests were significant and for $P(X = 1) = 0.8$ a rejection frequency of 6.12 % was observed for the elaborated multi-group model.

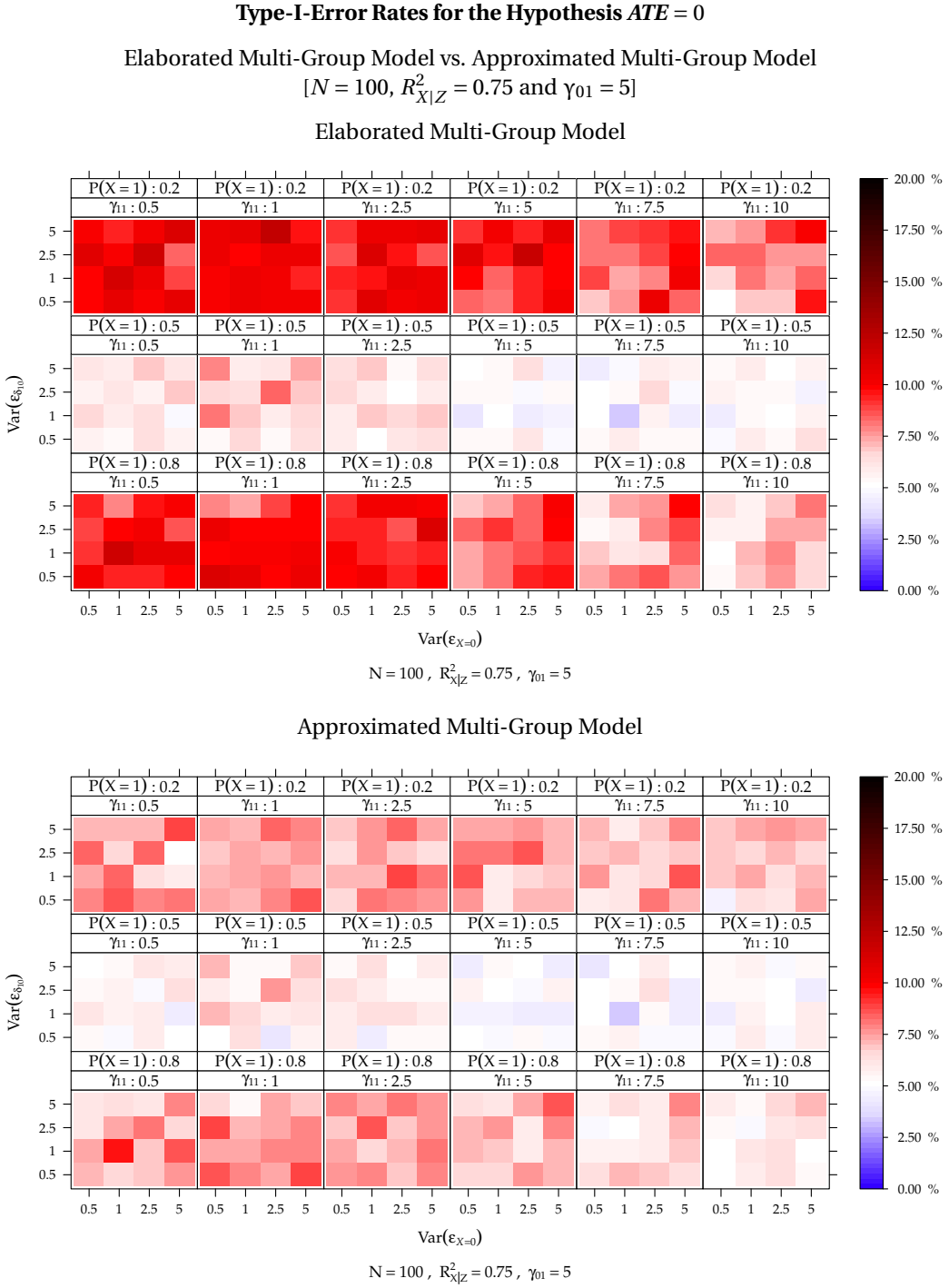


Figure 4.49: Type-I-error rate: Level plots for the elaborated multi-group model and the approximated multi-group model [$N = 100$, $R^2_{X|Z} = 0.75$ and $\gamma_{01} = 5$]

tion frequencies over all simulated conditions of part I of the simulation study (conditional on group size and sample size, see Figure 4.48) are acceptable at around the nominal 5 % level and very similar to those presented for the elaborated multi-group model (see Figure 4.42). For medium and large sample sizes

($N >= 250$) or for equal group sizes we obtained rejection frequencies for tests of the hypothesis $ATE = 0$ based on the approximated multi-group model within the corresponding confidence intervals.

To justify the small sample comparisons performed in the second Monte Carlo simulation (see section 4.8 for the results of simulation study II), Figure 4.49 compares the empirical type-I-error rates between the elaborated multi-group model and the approximated multi-group model for $N = 100$.[68] Furthermore, Table 4.10 presents the empirical type-I-error rates averaged over all conditions of simulation study I, grouped by sample size N and group size $P(X = 1)$. Obviously, the approximated multi-group model not only yields results comparable to the elaborated multi-group model, the augmentation-based approach in fact outperforms the `knownclass`–approach for datasets with unequal group sizes.[69] The advantage of the approximated multi-group model is larger for datasets with a strong dependency between X and Z (see Table 4.11).

Table 4.12: Relative bias of the standard error of the *ATE*–estimator: Approximated multi-group model and the elaborated multi-group model, grouped by dependency between X and Z $R^2_{X|Z}$, sample size N and group size $P(X = 1)$

			Approximated Multi-Group Model			Elaborated Multi-Group Model		
N	$R^2_{X\|Z}$	$P(X=1)$	0.5	0.2	0.8	0.5	0.2	0.8
100	0.75		−1.75 %	**−7.46** %	**−6.24** %	−2.22 %	**−11.52** %	**−9.94** %
250			−0.86 %	−2.44 %	−2.55 %	−1.02 %	−4.64 %	−3.84 %
400			−0.2 %	−1.56 %	−1.3 %	−0.43 %	−2.81 %	−2.49 %
1000			0.17 %	−0.64 %	−0.7 %	−0.48 %	−1.05 %	−0.95 %
100	0.5		−1.43 %	**−6.51** %	−4.96 %	−2.05 %	**−10.5** %	**−8.29** %
250			−0.44 %	−2.42 %	−2.04 %	−0.87 %	−3.87 %	−3.19 %
400			−0.31 %	−1.57 %	−0.92 %	−0.25 %	−2.24 %	−1.72 %
1000			−0.3 %	−0.72 %	−0.38 %	−0.28 %	−1.18 %	−0.74 %
100	0.25		−1.18 %	**−5.96** %	−4.44 %	−1.71 %	**−7.99** %	**−6.25** %
250			−0.72 %	−2.11 %	−1.8 %	−0.83 %	−3.06 %	−2.26 %
400			−0.12 %	−1.43 %	−0.96 %	−0.51 %	−1.62 %	−1.33 %
1000			−0.12 %	−0.2 %	−0.11 %	0.05 %	−0.73 %	−0.66 %
100	0.1		−1.45 %	−4.44 %	−3.02 %	−1.08 %	**−5.14** %	−3.63 %
250			−0.37 %	−1.54 %	−1.02 %	−0.49 %	−1.98 %	−1.38 %
400			−0.46 %	−0.69 %	−0.59 %	−0.21 %	−1.38 %	−0.87 %
1000			−0.25 %	−0.02 %	−0.1 %	0.11 %	−0.39 %	−0.07 %

Bias of the Standard Error of the *ATE*–Estimate The observed pattern of the relative bias of the standard error of the average total effect estimator for the approximated multi-group model is almost comparable to the pattern described for the elaborated multi-group model for large sample sizes (see Table 4.12).[70] These findings are in line with the results presented previously for the empirical type-I-error rates. The interesting differences between the elaborated multi-group model and the approximated multi-group model for small

[68] Level plots for the empirical type-I-error rate obtained for larger sample sizes are printed in the digital appendix, see the additional Figure 87 on page 98 and the additional Figure 88 on page 99 of the digital appendix.

[69] Note that although the inflation of empirical type-I-error rates observed for the approximated multi-group model is generally smaller, the observed rejection frequencies for simulated conditions with small sample sizes and unequal group sizes are still above the nominal level for datasets generated with small and medium interaction effects ($\gamma_{11} < 5$, see Figure 4.49).

[70] See also the additional Figure 89 on page 100 and Figure 90 on page 101 of the digital appendix.

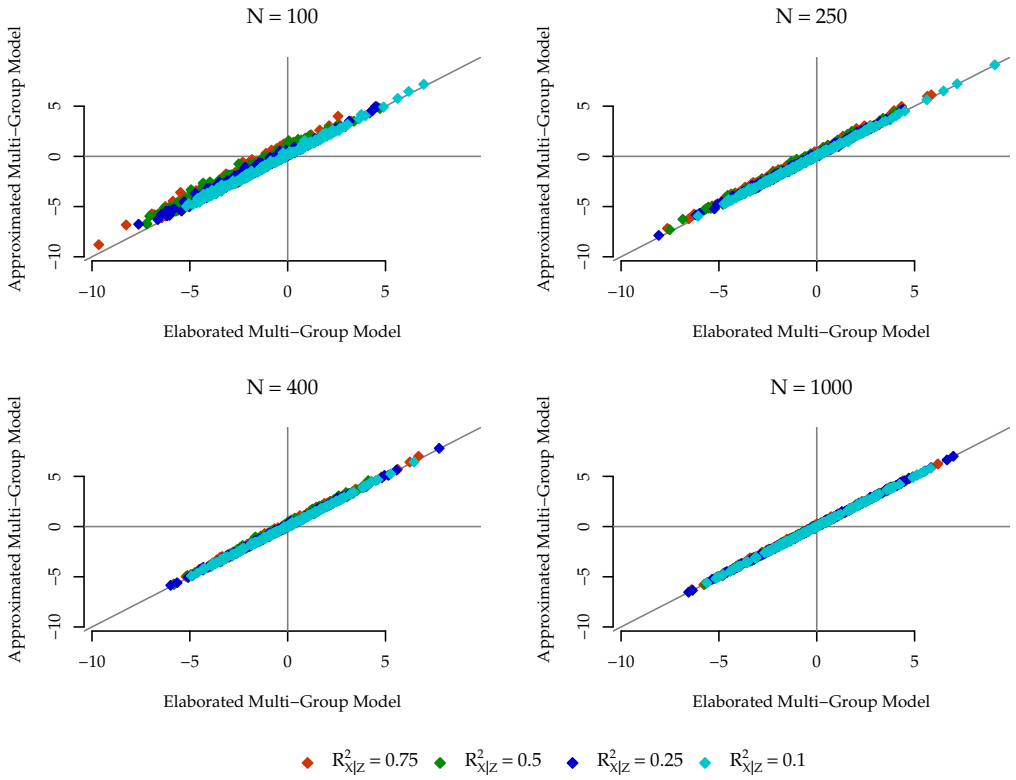


Figure 4.50: Relative bias of the standard error of the *ATE*–estimator: Scatter plots for the elaborated multi-group model vs. the approximated multi-group model, grouped by sample size N [$P(X = 1) = 0.5$]

samples and strong dependencies of X and Z, as obtained from this part of the simulation study, are summarized in Figure 4.50.[71] For equal group sizes some systematic differences in favor of the approximated multi-group model can be observed. Averaged over all 768 cells of simulation study I with $N = 100$ and $P(X = 1) = 0.5$, the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ of the *ATE*–estimator is -1.76 % for the elaborated multi-group model and -1.45 % for the approximated multi-group model. For unequal group sizes, these differences become greater, in particular for small sample sizes. For treatment groups larger (smaller) than control groups, the bias of the standard error obtained from the elaborated multi-group model is on average -8.79 % (-7.03 %), whereas the average total effect estimator for the approximated multi-group model for the same conditions is less biased -6.09 % (-4.67 %).[72]

[71] See also the additional Figure 91 on page 102 and Figure 92 on page 103 of the digital appendix.

[72] As the additional Figure 91 on page 102 and Figure 92 on page 103 of the digital appendix suggest, the relative bias of the standard error of *ATE*–estimator differs particularly for datasets, where X and Z are strongly connected (i. e., for $R^2_{X|Z} = 0.75$ and $R^2_{X|Z} = 0.5$), whereas for conditions with only a weak dependence of X and Z both methods yield comparable biases (see also table 4.12).

Summary The elaborated and the approximated multi-group models gave unbiased estimators of the average total effect with unbiased standard errors for large sample sizes. The standard error for the *ATE*–estimator based on the approximated multi-group model was found to be less biased for datasets with small sample sizes. Taking into account the empirical evidence that the asymptotic covariances between parameter estimates of the regression Y on Z and X and the estimated mean of the treatment variable X (used as estimate of the treatment probability) are uncorrelated, the approximated multi-group model becomes the most reasonable implementation for tests of hypothesis about the average total effect in generalized analysis of covariance. This preliminary conclusion is mainly based on the analysis of the absolute biases of the estimated average total effects, the relative biases of the standard error of the average total effect estimator, and based on the empirical type-I-error rates for the $ATE = 0$ hypothesis.

4.7.4 Summary

This section presented the results of our three correctly specified implementations of generalized analysis of covariance within the framework of structural equation modeling. We observed identical and unbiased estimators of the average total effect for the elaborated multi-group model and the approximated multi-group model and perfect convergence rates for all multi-group models. The performance of the elaborated single group model was not satisfying in the light of the obtained results for the elaborated multi-group model and in particular when compared to the results of the approximated multi-group model.

In the next section we will present the results of the second part of the simulation study, conducted to compare the small sample behavior and statistical power of the developed structural equation models. Although we will include the results obtained for the single-group implementations of generalized analysis of covariance, special focus will lie on a comparison of the statistical power of the elaborated and approximated multi-group models with the observed statistical power of the approach based on adjusted standard error for the average total effect estimated as a regression estimate.

4.8 Model Comparison

In this last section of the chapter we will compare the different implementations of generalized analysis of covariance with respect to both the sample sizes necessary to obtain appropriate type-I-error rates and the statistical power to detect average total effects. Furthermore, we will focus on the empirical convergence rates for the different structural equation models, especially for those conditions where the data are generated with true average total effects different from zero and for conditions with reasonably small sample sizes.

It is important to note that we studied the different implementations of generalized analysis of covariance in the second Monte Carlo simulation (part II) under less extreme conditions (see tables in section 4.2 for the selected parameters used in both parts of the simulation study). The underlying logic for the choice of the parameters used for generating the data was the following: Using the first part of the simulation study, potential drawbacks of the developed models under a wide range of possible conditions should be identified. The second part of the simulation study was designed to compare the different feasible implementations of generalized analysis of covariance under more realistic parameter sets. In particular compared to part I of the simulation study, we shall report averaged empirical rejection frequencies of the hypothesis $ATE = 0$ in the next subsection which validly reflect the expected performance of the data analysis techniques for empirical applications, for instance, under conditions with mild between-group residual variance heterogeneity.

4.8.1 Small Sample Behavior of the Adjustment Methods

In simulation study II, the following levels of the factor *sample sizes* were varied in order to study the small sample behavior of the possible implementations of a generalized analysis of covariance (see in Table 4.1 on page 106): $N = \{20, 30, 50, 75, 100, 150, 200, 250, 500, 1000\}$. Line charts are presented in this subsection to give a compact view of the observed small sample behavior with respect to the empirical type-I-error rates. Figure 4.51 [simple multi-group model (sample), simple single group (interaction) model, and elaborated single group model] and Figure 4.52 [elaborated multi-group model, approximated multi-group model, and regression estimates (normal approximation)] give the empirical type-I-error rates for all conditions of simulation study II where data were generated with a true average total effect of zero ($ATE = 0$, that means $d = 0$). Different charts are included in each figure, split by group size (equal group sizes in the second row of the 3 times 3 charts, unequal group sizes in the first and the third row). The observed rejection frequencies are visualized as line charts (gray lines) for the remaining 72 cells of the simulation study, in which each line connects the observed 10 rejection frequencies (on the y–axis) given the sample size (on the x–axis). Black lines with additional symbols for each sample size present the average over all cells in the plot, again conditional on the sample size.

The horizontal straight lines within the figures 4.51 and 4.52 mark the confidence intervals for the rejection frequencies (i. e., the acceptable empirical type-I-error rates) as expected for data generated without a true average total effect and for $N_{\text{Rep}} = 1000$ replications. Additionally, the meaningful points where

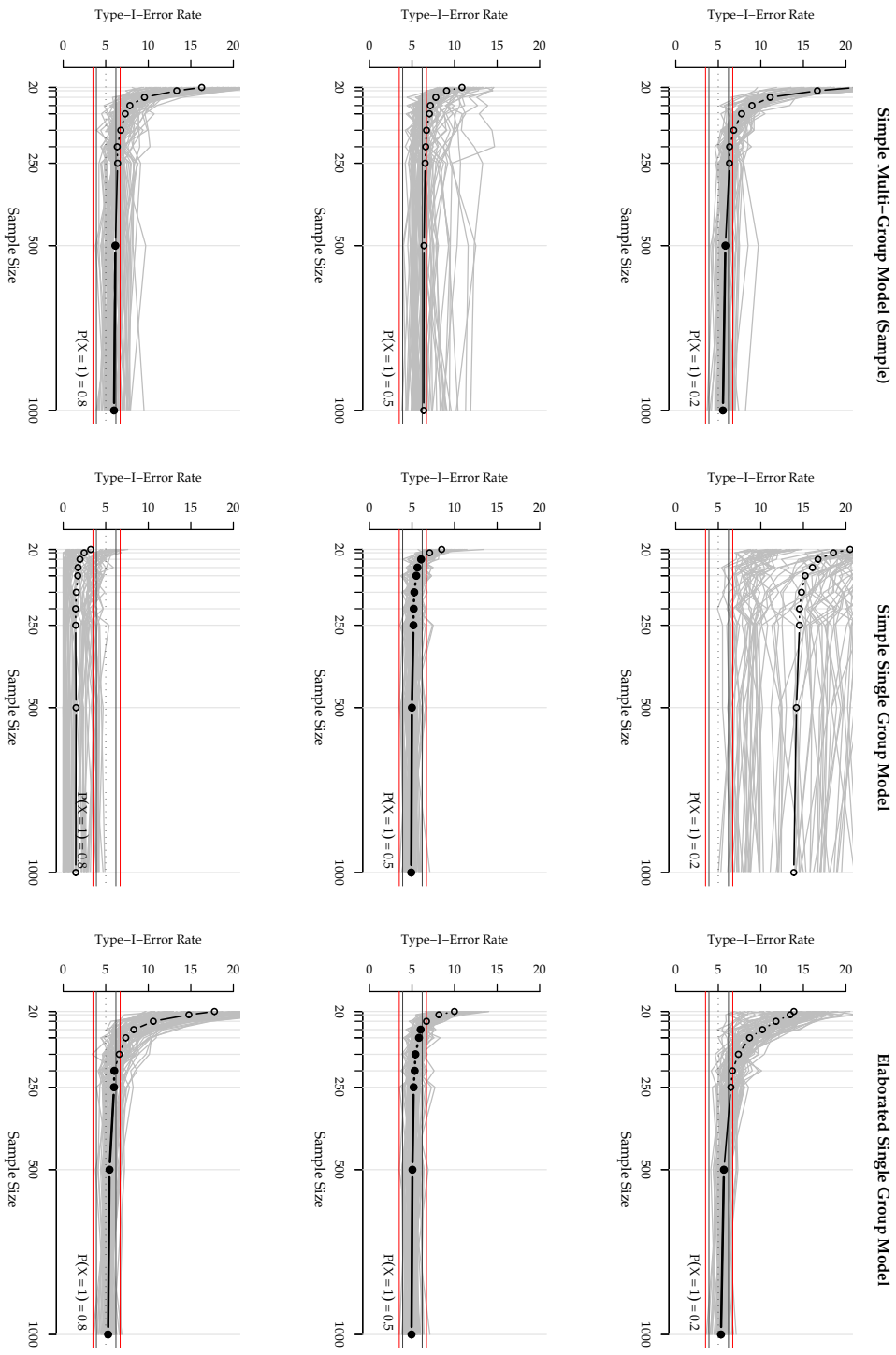


Figure 4.51: Type-I-error rate: Line plots for the simple multi-group model (sample), the simple single group model (with interaction) and the elaborated single group model, conditional on sample size N (simulation study II) [$d = 0$]

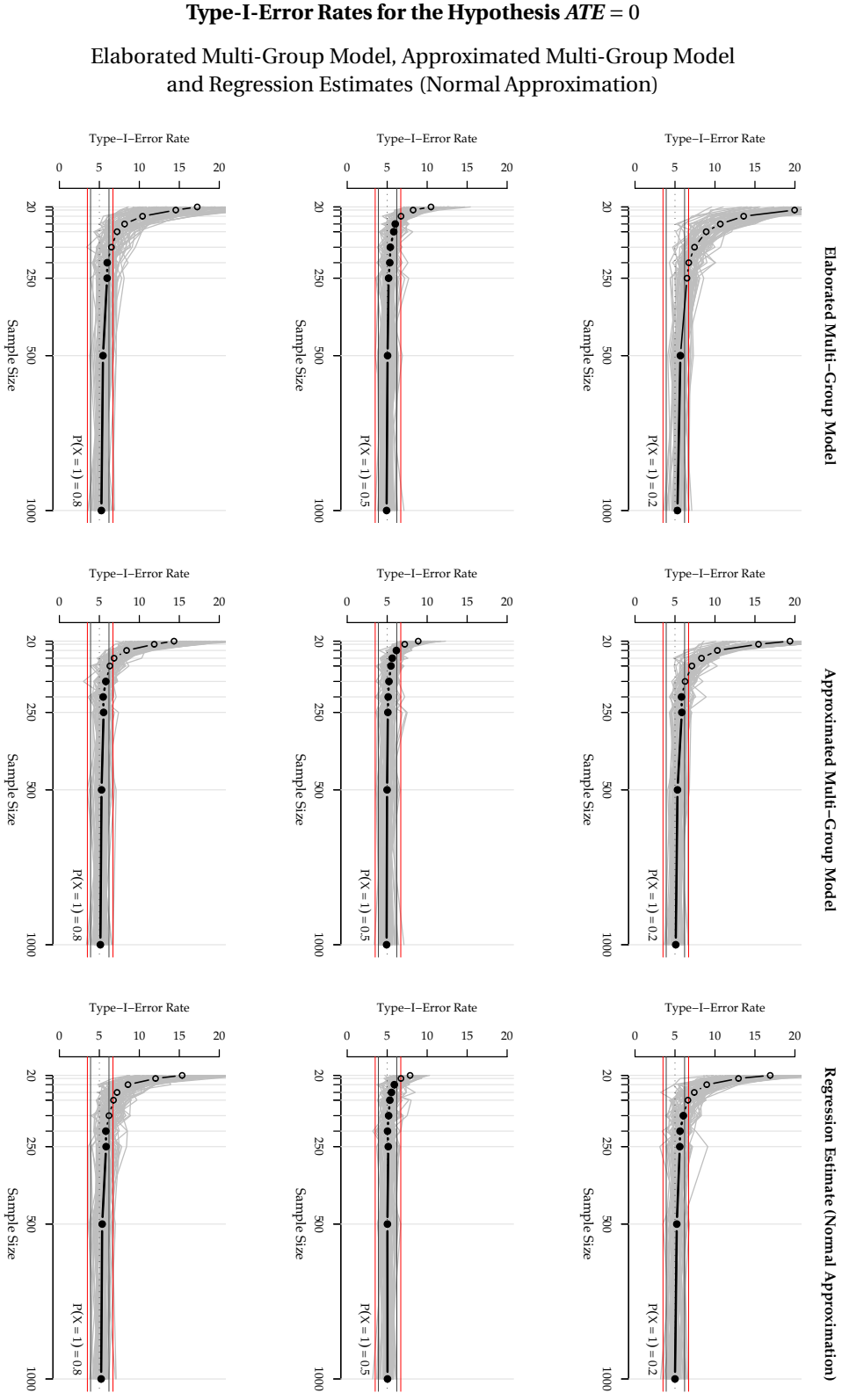


Figure 4.52: Type-I-error rate: Line plots for the elaborated multi-group model, the approximated multi-group model and regression estimates (normal approximation), conditional on sample size N (simulation study II) [$d = 0$]

the average[73] of the rejection frequencies is within the 99 % confidence intervals are highlighted with solid circles. For instance, the chart for the small sample performance of the simple multi-group model (sample) for unequal group sizes with $P(X = 1) = 0.2$ (presented in the first row and the first column of Figure 4.51) can be read in the following way: The observed empirical type-I-error rates are within the expected confidence interval for sample sizes of $N = 500$ and $N = 1000$. This is indicated by the two solid circles. Furthermore, the plot shows that for $N = 1000$ there are still several conditions that result in empirical type-I-error rates above the upper limit of the 99 % confidence interval. The second chart in this column presents the observed rejection frequencies for the simple multi-group model (sample) for equal group sizes with equal treatment probabilities $P(X = 1) = 0.5$. Obviously, even for the largest sample size included in the simulation study, the averaged type-I-error rate is outside of the confidence interval. An inspection of the second column of Figure 4.51 repeats the findings presented before for the simple single group in subsection 4.6.2 and subsection 4.7.1: The observed rejection frequencies do not converge to the nominal type-I-error rate for unequal group sizes for any of the included sample sizes (first and third row of the second column in Figure 4.51).

Table 4.13 presents the rejection frequencies for the hypothesis $ATE = 0$ as well as the relative bias of the standard error of the *ATE*–estimator for all conditions of simulation study II with no average total effect ($d = 0$ and $ATE = 0$). The rejection frequencies for all methods included in this table converge to the nominal 5 % level. For the elaborated multi-group model, the approximated multi-group model, and the regression estimate approach the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ consistently decreases with increasing sample sizes. Furthermore, note that the standard errors obtained from the approximated multi-group model are less biased compared to those computed for the $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ based on the regression estimates in almost all conditions.

Only the sequence of standard errors observed for the *ATE*–estimator based on the elaborated single group model and sample sizes smaller than $N = 75$ does not decrease systematically.[74] These irregularities can be explained by the observed convergence rates of the elaborated single group model.[75] The elaborated single group model has substantive convergence problems for all conditions with small and very small sample sizes, regardless of the treatment probability.[76]

[73] That is the averaged rejection frequency conditional on the sample size, computed across all cells of the simulation study with the appropriate treatment probability and with no average total effect, $d = 0$.

[74] The same behavior can be observed for conditions of simulation study II with $d = 0.2$ (see the additional table 9 on page 129 of the digital appendix), with $d = 0.5$ (see the additional table 10 on page 130 of the digital appendix) and with $d = 0.8$ (see the additional table 11 on page 131 of the digital appendix).

[75] See the additional table 12 presented on page 132 of the digital appendix.

[76] We present the observed convergence rates as additional figures in the digital appendix to show that the substantial convergence problems of the elaborated single group model are independent of the effect size (see the additional Figure 93 on page 104 of the digital appendix, and see also the additional Figure 94 on page 105, Figure 95 on page 106, and Figure 96 on page 107 of the digital appendix for the convergence rates observed for the remaining structural equation models).

Note that we observed no further problems in the different structural equation models with respect to the convergence rates. This was expected because the estimation of a restricted model is not required for testing the hypothesis of no average total effect with the help of the Wald–test (see subsection 3.3.2 for details), and because of the less extreme conditions analyzed in part II of the simulation study (see above).

Table 4.13: Rejection frequency and relative bias of the standard error for the *ATE*–estimator, grouped by sample size N and treatment probability $P(X=1)$ $[d=0]$

Sample Size	Elaborated Single Group Model		Elaborated Multi-Group Model		Approximated Multi-Group Model		Regression Estimate (t–test)	
				$P(X=1)=0.5$				
20	9.98 %	(**11.35** %)	10.66 %	(**−15.65** %)	9.04 %	(**−12.74** %)	6.94 %	(**−15.10** %)
30	8.16 %	(**−6.56** %)	7.95 %	(**−7.30** %)	6.96 %	(**−5.47** %)	6.15 %	(**−8.08** %)
50	6.70 %	(−3.59 %)	6.80 %	(−4.68 %)	6.25 %	(−3.68 %)	5.54 %	(−4.69 %)
75	6.01 %	(−2.47 %)	6.08 %	(−2.91 %)	5.69 %	(−2.24 %)	5.35 %	(−2.75 %)
100	5.82 %	(−1.85 %)	5.71 %	(−1.85 %)	5.50 %	(−1.31 %)	5.18 %	(−1.83 %)
150	5.40 %	(−0.97 %)	5.35 %	(−1.29 %)	5.22 %	(−0.96 %)	5.08 %	(−1.39 %)
200	5.31 %	(−0.23 %)	5.40 %	(−0.64 %)	5.25 %	(−0.37 %)	4.96 %	(−0.77 %)
250	5.19 %	(−0.52 %)	5.20 %	(−0.46 %)	5.11 %	(−0.27 %)	5.06 %	(−0.95 %)
500	5.05 %	(−0.09 %)	5.03 %	(0.25 %)	4.98 %	(0.36 %)	5.01 %	(−0.26 %)
1000	4.94 %	(0.41 %)	5.11 %	(−0.32 %)	5.12 %	(−0.27 %)	5.03 %	(−0.58 %)
				$P(X=1)=0.2$				
20	13.89 %	(−0.25 %)	24.66 %	(**−52.02** %)	19.52 %	(**−45.12** %)	15.97 %	(**−87.53** %)
30	13.45 %	(−1.60 %)	20.22 %	(**−43.21** %)	15.71 %	(**−36.91** %)	12.33 %	(**−66.76** %)
50	11.78 %	(**9.82** %)	13.63 %	(**−24.27** %)	10.31 %	(**−18.70** %)	8.63 %	(**−25.28** %)
75	10.19 %	(**−9.79** %)	10.38 %	(**−14.06** %)	8.14 %	(**−9.94** %)	7.22 %	(**−13.38** %)
100	8.67 %	(**−8.16** %)	8.84 %	(**−9.88** %)	7.15 %	(**−6.62** %)	6.43 %	(**−9.63** %)
150	7.36 %	(**−5.46** %)	7.43 %	(**−6.09** %)	6.22 %	(−3.93 %)	5.92 %	(**−5.98** %)
200	6.66 %	(−3.83 %)	7.00 %	(−4.86 %)	6.05 %	(−3.15 %)	5.55 %	(−4.62 %)
250	6.48 %	(−3.37 %)	6.39 %	(−3.27 %)	5.63 %	(−1.88 %)	5.54 %	(−3.62 %)
500	5.66 %	(−1.39 %)	5.65 %	(−1.72 %)	5.26 %	(−0.97 %)	5.19 %	(−1.92 %)
1000	5.32 %	(−0.40 %)	5.47 %	(−1.22 %)	5.25 %	(−0.83 %)	4.99 %	(−0.90 %)
				$P(X=1)=0.8$				
20	17.78 %	(−3.83 %)	17.19 %	(**−42.35** %)	14.26 %	(**−36.26** %)	14.30 %	(**−81.99** %)
30	14.77 %	(**−5.91** %)	14.36 %	(**−32.92** %)	11.60 %	(**−27.54** %)	11.36 %	(**−60.44** %)
50	10.59 %	(−1.08 %)	10.19 %	(**−15.90** %)	8.31 %	(**−11.91** %)	8.21 %	(**−20.18** %)
75	8.30 %	(**−5.24** %)	8.26 %	(**−9.47** %)	6.97 %	(**−6.66** %)	6.96 %	(**−9.09** %)
100	7.36 %	(**−5.63** %)	7.23 %	(**−6.43** %)	6.20 %	(−4.28 %)	6.60 %	(**−6.65** %)
150	6.58 %	(−3.67 %)	6.74 %	(−4.86 %)	6.06 %	(−3.47 %)	6.08 %	(−4.07 %)
200	6.01 %	(−2.54 %)	6.14 %	(−3.32 %)	5.62 %	(−2.22 %)	5.72 %	(−2.95 %)
250	5.98 %	(−2.38 %)	5.82 %	(−2.19 %)	5.40 %	(−1.29 %)	5.78 %	(−2.50 %)
500	5.44 %	(−1.30 %)	5.54 %	(−1.52 %)	5.33 %	(−1.07 %)	5.32 %	(−1.22 %)
1000	5.28 %	(−0.83 %)	5.21 %	(−0.56 %)	5.13 %	(−0.30 %)	5.21 %	(−0.82 %)

Note: Relative bias $RB\left[\widehat{S.E.}(\widehat{\text{ATE}}_{10})\right]$ *in parenthesis. Printed are averages over* 72 *cells of the simulation study.*

Elaborated Single Group vs. Elaborated Multi-Group The elaborated single group model and the elaborated multi-group model are comparable with respect to the rejection frequencies observed for small samples, as presented in Figure 4.51 and Figure 4.52: The obtained empirical type-I-error rates are within the

confidence intervals for equal group sizes and sample sizes of $N >= 75$. For unequal treatment probabilities with treatment groups smaller than the control groups, a minimum sample size of $N = 500$ was necessary to obtain rejection frequencies within the expected range. For unequal group sizes with treatment groups larger than the control groups a minimum sample size of $N = 200$ can be observed as the lower limit.

In section 4.7.1 we pointed out that whenever the elaborated single group model with random slopes is over-parameterized with respect to the residual variance of the random slope, the elaborated multi-group model and the approximated multi-group models are still feasible alternatives. Although the elaborated single group model does not stand out with inflated empirical type-I-error rates under the less extreme conditions of simulation study II, the small sample comparison gives no further indication of possible disadvantages of the two feasible implementations of generalized analysis of covariances as multi-group structural equation model with nonlinear constraints.

Approximated Multi-Group Model vs. Regression Estimate The most unexpected result of the first part of the Monte Carlo simulation was the excellent performance of the test statistic for the hypothesis of $ATE = 0$ based on the adjusted standard errors for the regression estimates. The small sample behavior of this test statistic with respect to the empirical type-I-error rates is plotted next to the results obtained for the approximated multi-group model in Figure 4.52. Both methods are absolutely comparable for equal group sizes in the sense that, when averaged across all conditions of simulation study II (with $d = 0$), a sample size of $N = 50$ is sufficient for a satisfactory type-I-error rate at the nominal α–level for both methods.

The approximated multi-group model is slightly superior compared to the regression estimate approach for unequal group sizes with $P(X = 1) = 0.8$ (the minimum sample size is $N = 150$ for the approximated multi-group model vs. $N = 200$ for the regression estimate approach). This advantage cannot be observed for the reverse conditions with treatment probabilities of $P(X = 1) = 0.2$. Here, the regression estimate approach achieves an averaged rejection frequency within the confidence bands for sample sizes larger than $N = 150$ (instead of $N = 200$ as the observed lower limit of the approximated multi-group model). Therefore, we conclude that the approximated multi-group model and the regression estimate are almost comparable with respect to the sample size requirements for a correct type-I-error rate.

Elaborated Multi-Group Model vs. Approximated Multi-Group Model The sample sizes necessary to obtain the nominal type-I-error rates for the elaborated multi-group model are consistently larger than the minimal sample sizes needed for the approximated multi-group model.[77] The observed differences be-

[77] To judge the importance of these small differences in more detail, the additional Figure 97 on page 108, Figure 98 on page 109 and Figure 99 on page 110 of the digital appendix compare the relative biases of the standard error of *ATE*–estimator obtained for the elaborated multi group model (x–axis) and the approximated multi-group model (y–axis). Within each figure different scatter plots are provided for each level of the factor sample size N as varied in the second part of the simulation study. Note that these scatter plots incorporate all conditions of simulation study II, i. e., also conditions with average total effects different from zero ($d \neq 0$) are

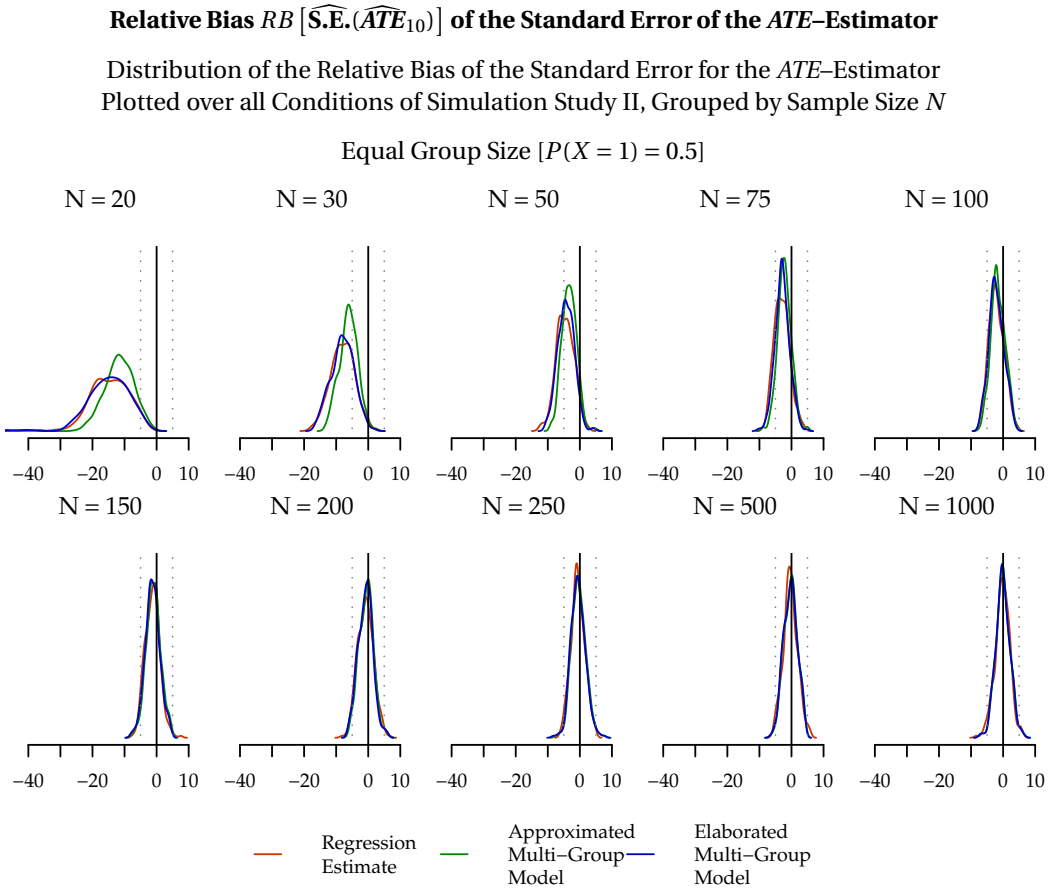


Figure 4.53: Relative bias of the standard error of the *ATE*–estimator: Distribution of the relative bias of standard error for the *ATE*–estimator for regression estimates, the approximated multi-group model and the elaborated multi-group model, grouped by sample size N (simulation study II) [$P(X = 1) = 0.5$]

tween the approximated and the elaborated multi-group model are connected to the sample size as well as to the dependency of X and Z (i. e., to $R^2_{X|Z}$), and larger differences are obtained for simulated datasets with unequal group sizes. The empirical distributions of the relative bias of the standard error of the *ATE*–estimator for different sample sizes (Figure 4.53 and Figure 4.54) indicate that there are clear benefits of the approximated multi-group model compared to the regression estimate and the elaborated multi-group model for equal group sizes with small samples ($N < 75$) and for unequal group sizes with medium sample sizes ($N < 250$). Nevertheless, it should be noted that the standard errors are estimated with a small negative bias for small sample sizes for all methods. Finally, it is interesting to recognize that the small sample behavior of the regression estimate approach and the elaborated multi-group model are very similar with respect to the relative bias of the standard error of the *ATE*–estimator.[78]

considered. Dotted black horizontal and vertical lines refer to the average relative bias of the standard error of the *ATE*–estimator [conditional on $P(X = 1)$ and N, according to the chart].

[78] See the detailed plots of the empirical distributions of $RB\left[\widehat{\text{S.E.}}(\widehat{ATE}_{10})\right]$ for all methods in the additional Figure 100 on page 111 (elaborated multi-group model), Figure 101 on page 112 (approximated multi-group model), Figure 102 on page 113 (regression esti-

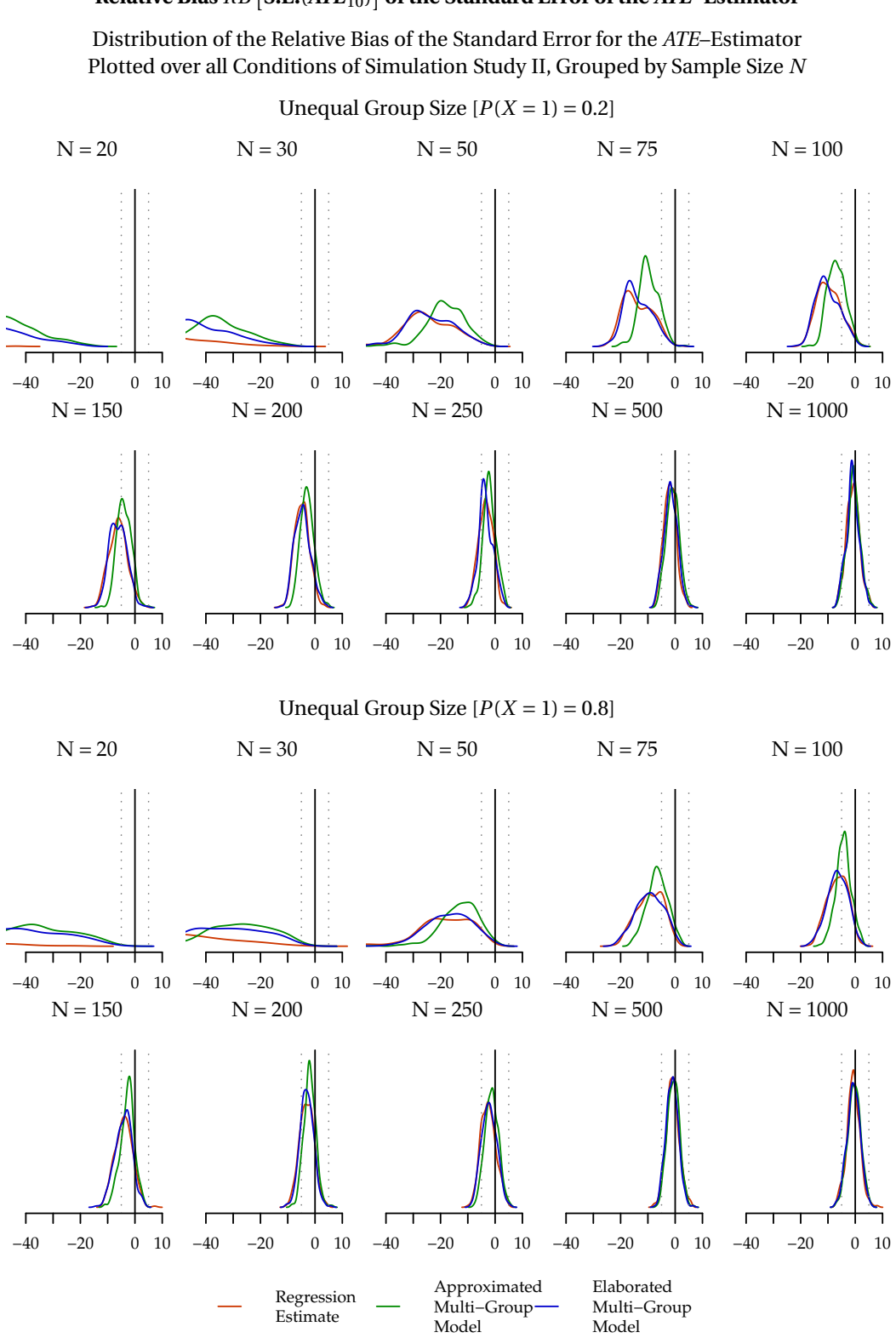


Figure 4.54: Relative bias of the standard error of the *ATE*–estimator: Distribution of the relative bias of standard error for the *ATE*–estimator for regression estimates, the approximated multi-group model and the elaborated multi-group model, grouped by sample size N (simulation study II) [$P(X = 1) = 0.2$ and $P(X = 1) = 0.8$]

mates), Figure 103 on page 114 (simple single group model), Figure 104 on page 115 (elaborated single group model), and Figure 105 on page 116 (simple multi-group model, sample) of the digital appendix.

4.8.2 Sample Size Requirements for Appropriate Statistical Power

The second part of the Monte Carlo simulation was designed to compare the different feasible implementations of generalized analysis of covariance with respect to the statistical power. For this reason, datasets with average total effects different from zero and with varying effect sizes ($d = 0.2$, $d = 0.5$, and $d = 0.8$) were generated and analyzed. Nevertheless, especially for small effect sizes, the conducted simulation study can give only a rough approximation to the intended model comparison because, due to computational burdens, only a few distinct levels of the factor sample size were simulated. In addition, the statistical power for the tests of hypotheses about the average total effect depends strongly on the amount of confounding (γ_{01}), and only two different values of this regression coefficient were used to generate the datasets. However, in total 6480 conditions with an average total effect different from zero were generated $N_{\text{Rep}} = 1000$ times and analyzed for the comparison of the statistical power of the final models (see section 4.3 for details) presented in this subsection.

The obtained rejection frequencies are plotted as line charts conditional on the sample size (see also the presentation of the necessary sample sizes for valid type-I-error rates given in subsection 4.8.1 for a description of the figures).[79] Within the two selected figures (see Figure 4.55 and Figure 4.56), there is a separate chart for each of the three different group size conditions (rows) and for the different statistical procedures (columns). The dotted horizontal line marks the rejection frequency of 80 % bordering the confidence interval as expected for $N_{\text{Rep}} = 1000$. In addition to the charts presented in Figure 4.51 and following, the lines are colored according to the value of the parameter γ_{01} ($\gamma_{01} = 1$ is printed in orange and $\gamma_{01} = 5$ is printed in blue). Therefore, the colored bold lines represent the average within the levels of γ_{01}, and the black bold lines show the overall average.

Figure 4.55 and Figure 4.56 are based on generated datasets with a small average total effect (effect size $d = 0.2$). Relative to the conventional criterion (rejection frequency of 80 %), the elaborated single group model, the approximated multi-group model, and the approach based on regression estimates yield comparable results. On average the criterion is reached for $\gamma_{01} = 5$ and for the following pairs: $N >= 500$ for $P(X = 1) = 0.2$, $N >= 200$ for $P(X = 1) = 0.5$, and $N >= 250$ for $P(X = 1) = 0.8$. The approach based on the adjusted standard error for the regression estimates produced a test statistic with a notably larger power compared to the developed structural equation models because this method reveals a small effect ($d = 0.2$) sufficiently more often for equal group sizes [e. g., for $\gamma_{01} = 5$ given a sample size of $N = 150$, and for unequal group sizes with $P(X = 1) = 0.8$ for $N = 200$].

[79] Even more detailed level plots of the statistical power for each cell in simulation study II are provided in the digital supplement (see Digital Supplement: 2-1).

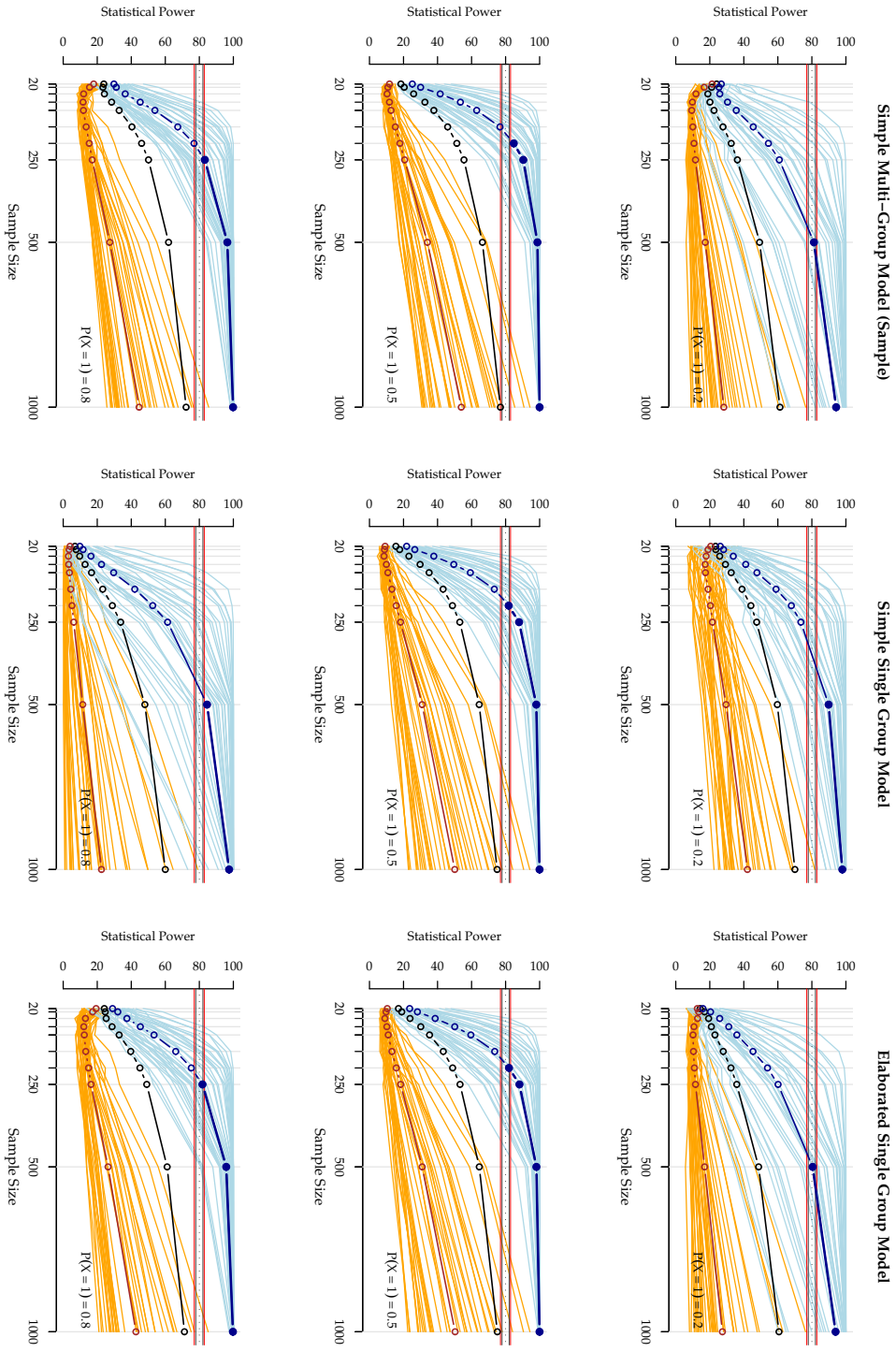


Figure 4.55: Statistical power to detect average total effects: Line charts for the simple multi-group model (sample), the simple single group model and the elaborated single group model, conditional on group size $P(X = 1)$ and sample size N [$d = 0.2$]

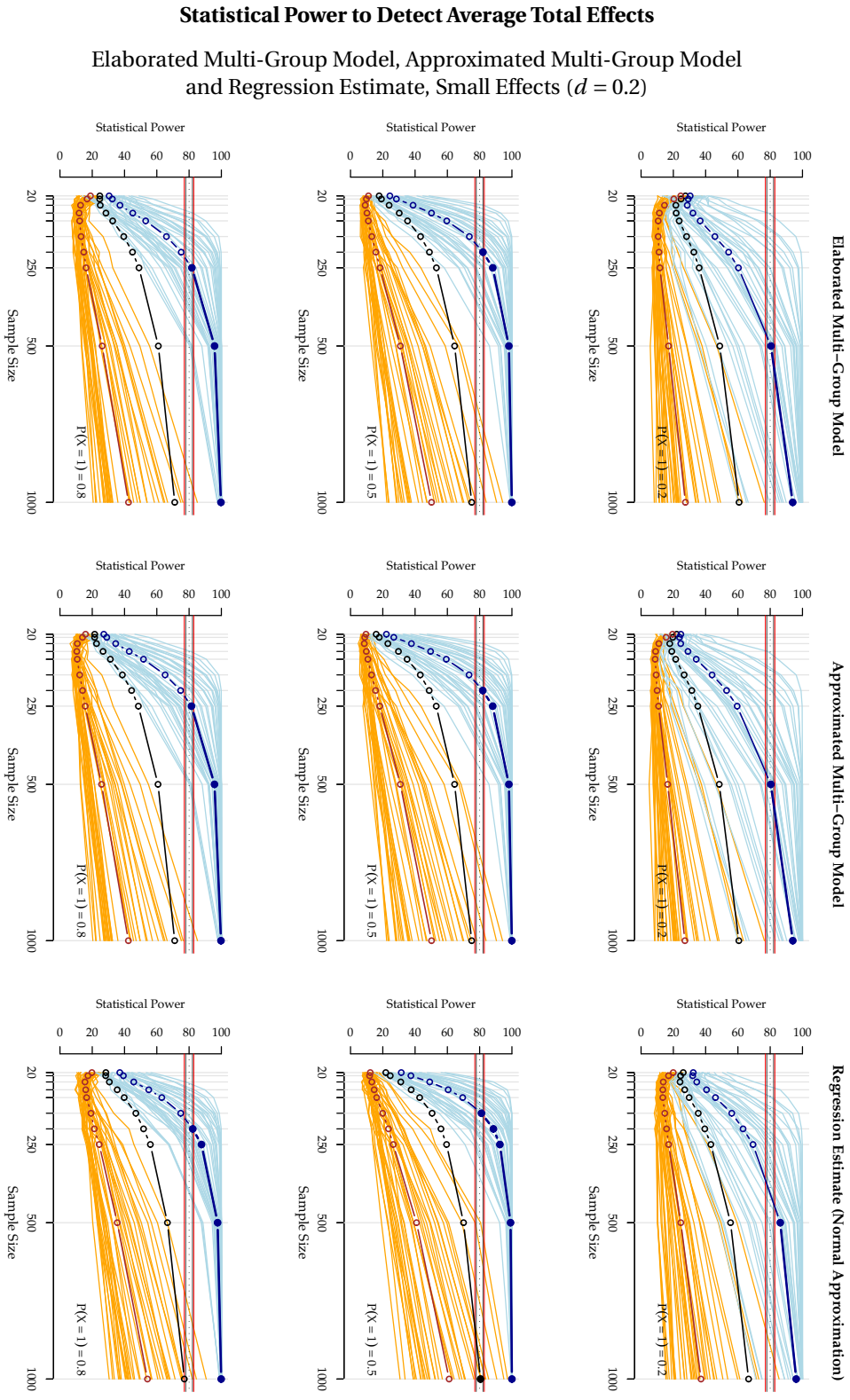


Figure 4.56: Statistical power to detect average total effects: Line charts for the elaborated multi-group model, the approximated multi-group model and the regression estimates, conditional on group size $P(X = 1)$ and sample size N [$d = 0.2$]

Similarly, small differences between the multi-group structural equation models and Schafer & Kangs' approach can be observed with respect to the statistical power for datasets with a medium average total effect (effect size $d = 0.5$).[80] The test based on the regression estimates achieves the criterion on average (i. e., regardless which value of γ_{01} was used for generating the datasets) with smaller sample sizes. For equal group sizes a minimal sample size of $N = 150$ was sufficient for the regression estimates (vs. $N = 200$ for the multi-group models), and for unequal group sizes (with $P(X = 1) = 0.8$) the criterion was reached for $N = 200$ (vs. $N = 250$ for the multi-group structural equation models).

Finally, there are several differences between the statistical power of the different implementations of generalized analysis of covariance for the datasets with a large effect size ($d = 0.8$).[81] For conditions with $\gamma_{01} = 1$ and $P(X = 1) = 0.2$, the test based on the nonlinear constraint of the approximated multi-group model as well as the Wald–test based on the elaborated single group model reject the hypothesis of no average total effect less often than does the test based on the elaborated multi-group model and the procedure based on the regression estimates. The regression estimates give acceptable rejection frequencies under these conditions for a sample size as low as $N = 150$.

Lastly, it should be noted that all studied methods result in a comparable statistical power according to the selected criterion for equal group sizes [$P(X = 1) = 0.5$].

[80] The corresponding line charts for medium effect size ($d = 0.5$) are included as addition Figure 106 on page 117 and Figure 107 on page 118 of the digital appendix.

[81] See the additional Figure 108 on page 119 and Figure 109 on page 120 of the digital appendix.

Chapter 5
Summary and General Discussion

In this thesis we studied statistical implementations of generalized analysis of covariance with nonlinear constraints in structural equation models (e. g., Nagengast, 2006; Flory, 2008; Steyer & Partchev, 2008), and developed these structural equation models further to adapt to requirements derived from the theory of stochastic causality (Steyer et al., in press). This chapter will summarize the presented theoretical considerations underlying the elaboration and evaluation of the implementations of generalized analysis of covariance and integrate the results of the Monte Carlo simulation into the summary and general discussion. Furthermore, we will critically discuss limitations of the investigated approach and the simulation study and we will highlight the need for further research into the enhancement and application of generalized analysis of covariance for the estimation of average total effects.

This chapter is organized in the following way: We will start with a summary of the theoretical considerations regarding generalized analysis of covariance which leads to the conclusion that the general linear model is not appropriate for the statistical implementation of this adjustment method. This will be followed by a synopsis of all partial results of the Monte Carlo studies, i. e., the findings obtained under conditions with almost homogenous between-group residual variances, the findings obtained under heterogeneity between-group of residual variances and the observations from the second simulation study with true average total effects different from zero. Afterwards, we will draw general conclusions for testing hypotheses about average total effects in quasi-experimental designs, i. e., we will sum up our analytical and empirical findings with respect to the adequacy of the studied implementations of generalized analysis of covariance as structural equation model with nonlinear constraints. Furthermore, we will review critical limitations of the presented research and describe interesting extensions as well as remaining research questions. Finally, we will present practical recommendations regarding the general linear hypothesis and the further development of the program package *EffectLite* (Steyer & Partchev, 2008).

5.1 Summary of the Theoretical Considerations

In this section we will start with the a summary of the considered requirements for the statistical implementation of generalized analysis of covariance and a brief comparison of this adjustment methods to alternative data analysis techniques for the estimation of average total effects. This will be followed by a synopsis of the presented discussion on unconditional inference about ordinary least-squares estimated average total

effects. Finally, we will review the structural equation models with nonlinear constraints and the research questions approached with our Monte Carlo simulation.

Theory of Causal Effects And Three Requirements for the Statistical Implementation We have started the thesis with a brief introduction of the theory of stochastic causality (Steyer et al., in press), as far as required to define the average total effect and to discuss the necessary assumptions for the identification of this average total effect by empirically estimable quantities for quasi-experimental designs. In contrast to the Rubin Causal Model (Rubin, 1974, 1977, 1978), the theory of stochastic causality is based upon a stochastic definition of the true outcomes. The first considered requirement for the statistical implementation of generalized analysis of covariance (heterogeneity of between-group residual variances) was derived from this conceptualization of the true outcome variables and the true effect variable as random variables.

Subsequently, we discussed causality conditions under which the average total effect can be identified by empirically estimable quantities, and we restricted the scope of this thesis to conditions where the assumption of Z–conditional unbiasedness of the covariate-treatment regression is fulfilled in non-randomized quasi-experimental designs. We emphasized that for the practical identification of the average total effect estimator the covariate-treatment regression is either specified with a parametric form or modeled non-parametrically and we described under which conditions a precisely correctly specified function form is necessary. Accordingly, the second requirement for the implementation of generalized analysis of covariance is the need to include covariate-treatment interactions because the exclusion of interaction terms would be an unnecessary strong restriction of the possible specifications of the covariate-treatment regression.

The theoretical considerations of causal effects and causal dependencies, e. g., the definition of the average total effect, refer to specific single-unit trials (Steyer et al., in press). In particular we described a basic single-unit trial for between-subject quasi-experiments as the scope of the structural equation models with nonlinear constraints studied in this thesis. Moreover, the third requirement considered for the implementation of generalized analysis of covariances (stochasticity of regressors, jointly distributed with the outcome variable) follows from the explicit notion of covariates and the treatment variable as random variables in this single-unit trial, and the conceptualization of the sample as generated by independent replications of the single-unit trial.

Finally, we presented the basic decomposition of the covariate-treatment regression into the intercept and effect functions (Wooldridge, 2001; Steyer et al., in press) and introduced the parametrization of generalized analysis of covariance used, for instance, in *EffectLite* (Steyer & Partchev, 2008).

Generalized Analysis of Covariance and Further Adjustment Methods In chapter 2, we reviewed different adjustment methods and discussed a within-study comparison, presented as an empirical example for the different data analysis techniques. We then utilized this brief compendium of adjustment methods to classify generalized analysis of covariance into the broader class of outcome modeling approaches and to clarify the relationship to the various alternative methods based on the propensity scores. We worked out that propensity score based adjustment methods are two-step approaches (Dehejia & Wahba, 1999) which often rely on a similar functional form assumption in applications as well as the adjustment methods based on the covariate-treatment regression (for instance, linearity and additivity of the logistic regression model used to estimate the propensity scores, Rosenbaum & Rubin, 1984). Moreover, we summarized that for most of the propensity score based adjustment methods the propensity scores are treated as known quantities (Imbens, 2004; Morgan & Todd, 2008), and that valid standard errors for applications with estimated propensity scores are still under research (Stuart, 2008).

Unfortunately, only a few studies deal with a comparison of the robustness of the outcome modeling approach in contrast to the robustness of the assignment modeling approach against misspecifications of the functional form assumption of the regression model (e. g., Drake, 1993). To circumvent this lack of clear evidence we utilized the empirical example to motivate the need for the development of a valid standard error for the estimated average total effect for generalized analysis of covariance. The empirical quasi-experimental example described in chapter 2 was organized as part of a larger experimental setting (Shadish et al., 2008a), resulting in a distinguishing design which facilitated the exemplary comparison of the reviewed adjustment methods in terms of the achieved absolute bias reduction of the adjusted *ATE*–estimates. In light of those within-study comparisons, the outcome modeling approaches, even with the common linear parameterization of the covariate-treatment regression, were found to be a serious alternative to the propensity score based methods.

Unconditional Inference About Ordinary Least-Squares Estimated the Average Total Effects Based on the assumptions of the general linear model we worked out that due to the covariate-treatment interaction(s) the expectation of the (possibly multivariate) covariate enters the *ATE*–estimator and accordingly invalidates the linearity assumption of the general linear hypothesis (Flory, 2004; Nagengast, 2006). We showed that the different existing methods for *probing interactions* (e. g., Rogosa, 1980; Hunka & Leighton, 1997; Cohen et al., 2003; Tate, 2004; Miyazaki & Maier, 2005; Bauer & Curran, 2005; Preacher et al., 2006), traditionally applied for moderated regressions (i. e., for covariate-treatment regressions with interaction terms), share the same weakness with respect to their validity for unconditional inferences, i. e., they are all based on the concept of inference conditional on fixed values of the covariate(s) [see also Rogosa, 1980]. We pointed out that in contrast to the conditional interpretation the average total effect has an uncondi-

tional meaning as the *net effect* of the treatment averaged over the distribution of the covariates (see, also Nagengast, 2009).

In order to draw valid unconditional inferences about average total effects, we revealed that the interplay between stochastic regressors and interaction terms invalidates the common simplification of the general linear model with respect to the fixed-X assumption (Allison, 1995; Chen, 2006), and showed that sampling properties of the regressors (i. e., the stochasticity of X and Z) must be considered for the statistical implementation of generalized analysis of covariance. We resolved the inconsistency between the literature on ordinary least-squares regression (i. e., the commonly claimed robustness of the ordinary least-squares regression to violations of the fixed-X assumption, e. g., Neter et al., 1983; Fisicaro & Tisak, 1994; Rencher & Schaalje, 2007) and the previous findings for generalized analysis of covariance (Flory, 2004; Nagengast, 2006; Flory, 2008). For valid inferences about average total effects, i. e., for unbiased standard errors of the *ATE*–estimator, the distinction between stochastic and fixed regressors is relevant (e. g., Sampson, 1974; Quinn & Keough, 2002) if covariate-treatment interactions are present. Additionally, we worked out that mean-centering (Aiken & West, 1996; Yang & Tsiatis, 2001; Wooldridge, 2001; West & Aiken, 2005) suffers from an identical mistreatment of the regressors' stochasticity (Cohen, 1978; Kromrey & Foster-Johnson, 1998; Imbens & Wooldridge, 2009).

Structural Equation Models with Nonlinear Constraints From the presented discussion of unconditional inference about the estimated average total effect it became clear that the statistical implementation of generalized analysis of covariance must take into account the joint distribution of covariates, the outcome variable and the treatment variable (Maddala, 1992). We focused on the well-known methodology of structural equation modeling, because a multivariate distributional assumption is typically made for the estimation of these models (see, e. g., Bollen, 1989), in order to develop a valid statistical inference for the *ATE*–estimator based on generalized analysis of covariance, which incorporates the mentioned requirements and additionally is flexible enough, for instance, to deal with missing values and to account for measurement error on the covariates.

Two conventional structural equation models were considered. The first model, the simple multigroup model (Sörbom, 1978) extended to covariate-treatment regressions with interaction terms (Nagengast, 2006; Flory, 2008; Steyer & Partchev, 2008), was shown to be misspecified with respect to the stochasticity of the treatment variable. The model parameters of this statistical implementation of generalized analysis of covariance are estimated by minimizing the group-specific distances between the observed and the implied variance-covariance matrices, without a distributional assumption for the treatment variable. If the mean of the treatment variable as fallible estimate of the group size is included and treated as a known constant in the nonlinear constraint, the variability of the average total effect estimator is underestimated

by the standard error obtained with the δ-method (Rao, 1973; Oehlert, 1992, 1992). The second model, the simple single group model as suggested by Flory (2008), was shown to be misspecified with respect to the implied variance structure. Due to the known effect of misspecifications on the parameter estimates for structural equation models, we excluded the simple single group model as a potential valid implementation of generalized analysis of covariance. Accordingly, only the robustness of the two conventional structural equation models against the described violations was formulated as a research question for the Monte Carlo simulation.

Furthermore, we developed two elaborated structural equation models as implementations of generalized analysis of covariance, while taking into account the reconciled and clarified misspecifications of the general linear hypothesis and the conventional structural equation models. Both elaborated structural equation models are correctly specified with respect to the discussed requirements and are expected to result in an unbiased *ATE*–estimator as well as in unbiased standard errors.

Using a detailed description of the application of the δ-method, we emphasized the importance of the asymptotic variance-covariance matrix of parameter estimates for the derivation of the standard error of the *ATE*–estimator (D. P. MacKinnon, 2008). We developed a strategy to judge the appropriateness of the underlying assumptions of the third valid implementation of generalized analysis of covariance, the approximated multi-group model (Nagengast, 2006; Steyer & Partchev, 2008), based on the maximum-likelihood estimation of model parameter for our elaborated multi-group model. Hence, we added an empirical verification of the assumption of uncorrelated parameter estimates, which is necessary for the augmentation approach upon which the approximated multi-group model rests, as a specific research question for the Monte Carlo study.

Model Comparison and Statistical Power The coincidence of the regressors' stochasticity and the need for covariate-treatment interactions played a crucial role in the discussion of the unconditional validity of the standard error of the *ATE*–estimator obtained from the general linear model. Heteroskedasticity presents a distinct challenge for the general linear model, but reasonable adjustments for standard errors are available using so-called robust standard errors (White, 1980a; J. G. MacKinnon & White, 1985; Zeileis, 2004; Greene, 2007). The empirical demonstration of the two different sources for standard error biases and a comparison of the small sample behavior of different robust estimators based on heteroskedasticity consistent variance-covariance matrices were formulated as specific research questions for the Monte Carlo simulation.

In contrast to the general linear model, the heterogeneity of residual variances as implied by the theory of stochastic causality was the distinguishing feature for the different structural equation models. Violations of the heteroskedasticity assumption for the general linear model result in biased standard errors for un-

equal group sizes (Berry, 1993; Hayes & Cai, 2007), whereas misspecified implied variances, as a particular kind of specification error of structural equation models, are known to result in biased parameter estimates (Kaplan, 1989; Curran et al., 1996; Yuan et al., 2003). Neither Nagengast (2006) nor Flory (2008) considered the heterogeneity of between-group residual variances for their simulation studies. Hence, we studied various degrees of heterogeneity of between-group residual variance in the Monte Carlo simulation to evaluate the robustness of the misspecified structural equation models.

A final research question was formulated to compare the correctly specified implementations of generalized analysis of covariance with respect to the small sample behavior as well as with respect to the statistical power. For this comparison, approached in the second part of the conduced Monte Carlo simulation, we also incorporated the adjusted standard errors for regression estimates (Schafer & Kang, 2008).

5.2 Summary of the Results of the Simulation Study

The detailed results obtained from both parts of the simulation study regarding the performance of all studied adjustment methods were presented in section 4.5 (general linear model), 4.6 (structural equation equation models under homogeneity of residual variances), 4.7 (structural equation equation models under heterogeneity of residual variances) and 4.8 (model comparison). In the following summary we will relate the particular results to the more global aims of the Monte Carlo simulation: the demonstration of consequences of the theoretical derivations, the analysis of the robustness of the misspecified structural equation models and a comparison of the performance of the theoretical suitable methods for testing average total effects.

5.2.1 Ignoring the Stochasticity of Covariate (Z) and Treatment Variable (X)

We demonstrated that the *ATE*–estimators based on ordinary least-squares estimated covariate-treatment-regressions are unbiased for all conditions of the presented simulation study, although the covariate was generated as a stochastic random variable whose realized values varied from (generated) sample to (generated) sample. This result was expected due to the unbiasedness property of the conditional estimated regression coefficients [see Equation (3.29) on page 68]. Accordingly, we referred to these *ATE*–estimates as the baseline for the comparison of the point estimates obtained from the multi-group structural equation models. Thus, we conclude that the average total effect estimator for the multi-group implementation of generalized analysis of covariance is unbiased for stochastic covariates and stochastic treatment variables and that the empirical variability of the *ATE*–estimates is equal to the variability of the ordinary least-squares estimates of the average total effect.

Stochasticity of Covariate (Z) To demonstrate consequences of the nonlinearity of the hypothesis of no average total effect we formulated the prediction that underestimated standard errors and therefore inflated type-I-error rates will be obtained from the general linear model (see the research question formulated in subsection 3.4.1). In line with our theoretical consideration of the unconditional variance of the *ATE*–estimator for covariate-treatment regressions with interaction terms and stochastic regressors, the Monte Carlo simulation impressively demonstrated these standard error biases. The observed (true) variability of the estimates for the average total effect was heavily underestimated by the standard error for the *ATE*–estimator from the general linear model for all datasets generated with non-zero interaction terms. Accordingly, ordinary least-squares estimated moderated regression provides no valid test statistic for the *ATE*–estimator if the slopes differ between groups (covariate-treatment interaction) and if the covariates are stochastic (observational studies). Exactly the same is true for the test statistics obtained for the simple regression coefficient when the covariate(s) are mean-centered with the estimated mean(s).

In particular, the general linear model is not robust against strong violations of the assumption of fixed regressors. The nonlinear dependency of the standard error bias from the amount of interaction was developed theoretically by working out the mathematical relationship between the unconditional variance of the estimator and the interaction parameter [see Equation (3.32) on page 69]. We confirmed this derivation based on simulated datasets, i. e., only slightly biased standard errors of the *ATE*–estimator were observed for small interaction terms and (nonlinear) increasing biases were observed for conditions with increasing interaction effects.

Stochasticity of Treatment Variable (X) In addition to the stochasticity of the covariate, the results of the simulation study demonstrate that it is necessary to consider the treatment variable as a random variable. Because the group size is treated as known number for the (conventional) multi-group structural equation models we predicted biased standard errors of the *ATE*–estimator and therefore inflated type-I-error rates for this implementation of generalized analysis of covariance (see the research question formulated in subsection 3.4.4). Confirming our theoretical considerations of the δ-method and the Wald–test, we found biased standard errors for the *ATE*–estimator from the (conventional) structural equation model ignoring the randomness of the treatment variable. However, the overall inflation of the type-I-errors rate had a smaller magnitude compared to the observed failure of the general linear model.

The consideration of the stochasticity of the treatment variable and, accordingly, the incorporation of the group size as an estimated parameter of the structural equation model was necessary for the development of a valid implementation of generalized analysis of covariance based on nonlinear constraints in structural equation models. From a statistical point of view, we provide strong evidence that the simple multi-group model gives inflated type-I-error rates if the covariate interacts with the treatment variable.

The observed bias of the standard error for the *ATE*–estimator did not vanish and accordingly the type-I-error rates did not converge to the nominal level even for large sample sizes.

5.2.2 Robustness to Heterogeneity of Residual Variances

The results of the simulation study confirmed the prediction derived from the literature that there are two distinct factors necessary to consider when implementing a valid test statistic for the hypothesis of no average total effect for data obtained from quasi-experimental designs: The stochasticity of the covariate on the one hand (as summarized in the previous subsection), and the heterogeneity of residual variances on the other. Nevertheless, different results for the general linear model and the single group structural equation models were obtained.

General Linear Model For all tests based on unadjusted standard errors for ordinary least-squares estimated average total effects, violations of the assumption of homoskedasticity gave biased standard errors for the *ATE*–estimator for unequal group sizes. To correct the ordinary-least square standard error for the effect of heteroskedasticity, we can generally endorse the application of heteroscedasticity consistent estimators for unequal group sizes based on our results, for instance, the `HC3` correction. Nevertheless, non of the studied robust standard errors was found to be unbiased for large interaction effects. To summarize these findings: The unadjusted ordinary least-squares standard error for the average total effect estimator is biased if the group-specific regressions are non-parallel or if the between-group residual variances are not equal, whereas the standard errors corrected with heteroscedasticity consistent estimators of the variance-covariance matrix are only susceptible to a negative bias due to the stochasticity of the regressors. Accordingly, the general linear model corrected for heteroskedasticity tends to inflate type-I-error rates only for tests of the hypothesis $ATE = 0$ under conditions with substantial covariate-treatment interactions.

Single Group Structural Equation Models with Nonlinear Constraints Similar to the unadjusted general linear model, the simple single group model was not found to be robust against violations of the implied structure of homogenous residual variances. For the single-group implementation of generalized analysis of covariance suggested by Flory (2008), we even observed absolute biases of the *ATE*–estimator for conditions with unequal group sizes and heterogeneity of between-group residual variances. This can be interpreted as empirical evidence that within the framework of structural equation modeling, misspecification of the estimated model can lead to seriously biased parameter estimates. Barring these absolute biases of the average total effect estimator under some selected conditions of the Monte Carlo simulation, we generally found the tests of the hypothesis $ATE = 0$ for the simple single group model to be misleading under variance heterogeneity for conditions with unequal group sizes due to two observations: heavily biased standard errors for the *ATE*–estimator and serious convergence problems.

Regrettably, we found the elaborated single group model to be prone to effects of a misspecified implied variance structure as well. Although no systematic biases of the average total effect estimator were observed, the standard errors, the resulting test statistics, and in addition, the obtained convergence rates depend partly on the appropriateness of the specified implied variance structure.

5.2.3 Asymptotic Variances and Covariance

We demonstrated that with respect to the accuracy of the estimated asymptotic variance-covariance matrices the standard errors obtained for the elaborated multi-group model and the approximated multi-group model are unbiased for sample sizes larger than $N = 100$ (see research question in subsection 3.4.3). This means that for generalized analysis of covariance implemented as appropriately specified multi-group structural equation model the standard errors of the *ATE*–estimator (as computed with the multivariate δ-method) are acceptable for sample sizes larger than $N = 50$ for each group. In general, the accuracy of asymptotic variance-covariance matrices was acceptable for all appropriately specified structural equation models. Hence, we replicated the finding of Nagengast (2006) and extended the scope of the available Monte Carlo evidence to conditions where the true average total effect differs from zero and to conditions with heterogeneity of between-group residual variance (see also Nagengast, 2009). The reported findings of the single group structural equation models — when the implied mean structure as well as the implied variance structure meet the structure produced by the selection of parameters for the data generation — can also be interpreted as an indication that the standard errors obtained from the full-information maximum likelihood `LMS`–estimator are appropriate for the analyzed nonlinear function of estimated model parameters. Beside the effect of the misspecified implied variance structure mentioned above, we observed trustable standard errors computed with the δ-method in our *statistical simulation* (see, e. g., D. P. MacKinnon, 2008).

With respect to the approximated multi-group approach (see the research question formulated in subsection 3.4.5) the results obtained in both parts of the Monte Carlo simulation support the assumption of partially uncorrelated parameter estimates, which has been formulated as a prerequisite for the augmentation approach (Nagengast, 2006, see also subsection 3.3.3.3). The investigated asymptotic covariances between the estimated group size (as an additional model parameter of the elaborated multi-group model) and the model parameters of the conventional multi-group structural equation models were found to be zero under all studied conditions (barring random fluctuations). In conjunction with the observed overall performance of the approximated multi-group model the results provide support for the augmentation-based implementation of a generalized analysis of covariance.

5.2.4 Regression Estimate and Predictive Simulation

The simulation-based procedure for the approximation of standard errors for *ATE*–estimates obtained from predicted scores (Gelman & Hill, 2007, see research question in subsection 3.4.6), result in negligible differences compared to the common general liner model. Although the average total effect estimator itself was found to be unbiased, the simulated standard errors were as much biased as the standard errors obtained from the mean-centering approach, both due to regressors' stochasticity and due to between-group residual variance heterogeneity.

In contrast, the adequateness of the adjusted standard error for the regression estimate approach (Schafer & Kang, 2008) was cogent under all conditions studied in the part I of the Monte Carlo simulation. Regardless of the amount of interaction and for all studied violations of the homoskedasticity assumption, the adjusted standard errors for the *ATE*–estimator were found to be unbiased for medium and large sample sizes. Accordingly, we compared the performance of the developed structural equation models with nonlinear constraints to the results obtained based on the adjusted standard errors for the regression estimates in the second part of the simulation study.

5.2.5 Sample Size Requirements and Model Comparison

A final comparison of the different feasible implementations of generalized analysis of covariances was performed with respect to the minimal sample size necessary to obtain correct type-I-error rates for datasets with a population level (true) average total effect of zero and with a special focus on the required sample size for an adequate statistical power for data generated with an average total effect of varied effect sizes (see research question in subsection 3.4.7).

The statistical power for the procedure based on adjusted standard errors for the ordinary least-squares estimated regression estimates was observed to be conspicuously higher due to a predominant small sample behavior. For tests of the hypothesis of no average total effect based on structural equation models with nonlinear constraints, the sample size requirements of the elaborated single group model were found to be less favorable compared to the multi-group models, especially for equal group sizes. Furthermore, a small advantage of the approximated multi-group model compared to the elaborated multi-group model was observed. Taking into account the finding that the asymptotic covariances between the model parameters of the approximated multi-group model and the estimated group size are estimated as zero, these differences might be described as the result of an over-parameterization of the elaborated multi-group model.

5.3 General Conclusions

How to implement generalized analysis of covariance? In conformity with Flory (2004) we conclude that tests of the hypotheses of no average total effect based on the general linear hypothesis result in inflated type-I-error rates. This conclusion is based on our analytical discussion of the variance of the *ATE*–estimators and again was demonstrated with results of our Monte Carlo simulation. Hence, a distinction between random and fixed regressors in regression models is relevant for testing hypotheses about average total effects in quasi-experimental designs. The covariate-treatment interactions invalidate the commonly used simplification of the random regression model for the implementation of generalized analysis of covariance (see subsection 5.4.1). All different test statistics based on ordinary least-squares estimates (including test statistics based on mean-centered covariates) are not capable of yielding valid implementations of generalized analysis of covariance (with interaction terms in the covariate-treatment regression), if the regressors are stochastic.

Generalized analysis of covariance based on multi-group structural equation models with nonlinear constraints and in particular the selected strategy for testing hypotheses on the average total effect, i. e., the Wald–test, performed reasonable well. This includes our developed elaborated multi-group model as well as the augmentation approach implemented in Steyer and Partchev (2008).

Moreover, another strategy to test hypotheses on average total effects for covariate-treatment regressions with interaction terms can be generally recommended. Our simulation study confirmed that a valid statistical implementation of generalized analysis of covariance, which incorporates the three considered requirements for the statistical model, can be obtained based on the recently published adjusted standard errors for regression estimates (Schafer & Kang, 2008).

Specification Error and Single Group Structural Equation Models The specification error of the implied variance structure within the framework of structural equation modeling was shown to have serious effects not only on the estimated standard errors of parameter estimates, but also on the parameter estimates themselves. This phenomenon was observed only for single group structural equation models. Hence, as the average total effect estimator is computed as a (nonlinear) function of the estimated parameters, caution is recommended with respect to the careful specification of the variance structure of the structural equation model used for generalized analysis of covariance. In general, we recommend neither the simple single group model developed by Flory (2008), because this gives clearly misleading results for heterogenous between-group residual variances, nor our newly developed elaborated single group model. If no further information about the amount of heterogeneity of residual variances is available for a real data application, the specification of the additional error variance for the random slope residual (which con-

stitutes the main difference between the simple and the elaborated single group model) has to be investigated empirically, for instance, with a likelihood ratio test for a comparison of the nested (single group) structural equation models. Nevertheless, the single group models might serve as a starting point for the development of dual modeling strategies which simultaneously incorporate (latent) regression models and the (logit transformed) propensity score (see section 5.4.2). Accordingly, with respect to the two general modeling strategies within the framework of structural equation modeling (single group modeling versus multi-group modeling), we conclude that multi-group approaches are generally more appropriate with respect to the implied variance structure (heterogeneity of residual variances). Both of the considered single group alternatives performed reasonably well for data generated fittingly to the specified (implied) variance structure. The multi-group models were observed to be less sensitive to an over-parameterization or under-parameterization of the implied structure of residual variances.

Regression Estimates vs. Structural Equation Models with Nonlinear-Constraints Empirically, the elaborated and the approximated multi-group implementations of generalized analysis of covariance in general performed comparable to the test statistic based on adjusted standard errors for regression estimates for medium and large sample sizes. Hence, with respect to the comparison of the implementations of generalized analysis of covariance, the following three unique features of the studied structural equation models should be emphasized.

Firstly, it is well known in the literature on structural equation modeling that the treatment of missing values is strongly facilitated by the available full-information maximum likelihood estimation methods (see, e. g., Wothke, 2000; Enders & Bandalos, 2001; Graham, 2003). At least as long as the missing data are *missing at random* (MAR, see, e. g., Rubin, 1976; Schafer, 1997; Schafer & Graham, 2002) the multi-group structural equation models estimated with the available program packages (for example `LISREL`, Jöreskog & Sörbom, 1996 - 2001, `EQS`, Bentler, 1995, and `Mplus`, L. K. Muthén & Muthén, 1998 - 2007) will result in unbiased *ATE*–estimators. Missing values of covariates typically need to be treated as an additional step for alternative adjustment methods, for instance, such as techniques based on estimated propensity scores (see, e. g., D'Agostino & Rubin, 2000; D'Agostino, Lang, Walkup, Morgan, & Karter, 2001, but also Rosenbaum, 2010, for an alternative approach). The preparation of the dataset, for example, by multiple imputation (see, e. g., Rubin, 1987), is necessary even for the regression estimates because the available adjusted standard errors are based on ordinary least-squares regressions.

Secondly, the developed structural equation models are well prepared to account for measurement error of the covariates. As described in section 2.5.5, this is an unique property of generalized analysis of covariance implemented as a structural equation model. It is expected that under conditions where Z-

conditional unbiasedness of the covariate-treatment regression holds and the covariate is a latent variable, an unbiased average total effect estimator will be obtained only by accounting for the measurement error.

Thirdly, the structural equation models can simply be extended to more than two treatment groups (see Steyer et al., in press, for two different hypotheses for more than two groups). According to our knowledge, however, the adjusted standard errors for regression estimates have not yet been developed for the simultaneous comparison of multiple groups.

The regression estimates approach based on ordinary least-squares estimated group-specific covariate regressions should serve as a benchmark for further developments and extensions of generalized analysis of covariance, i. e., for the newly developed structural equation models as well as for the already implemented augmentation approach (Nagengast, 2006; Steyer & Partchev, 2008). The results of both parts of the simulation study showed no evidence that the least-squares estimated regression estimation approach is in general superior to the maximum likelihood estimated structural equation models for data generated corresponding to the underlying multivariate normality assumption. To sum up, we prefer the implementation of generalized analysis of covariance as elaborated or approximated multi-group structural equation model with nonlinear constraints, basically because of the flexibility of the framework of structural equation modeling.

5.4 Limitations and Further Research

The presented discussion on the implementation of generalized analysis of covariance as a structural equation model with nonlinear constraints comes with several limitations and reveals open research questions. We will summarize limitations in the next subsection starting with a description of shortcomings of the conducted Monte Carlo simulations which can be easily approached in subsequent research. Additionally, we will describe the more general limitations of generalized analysis of covariance as studied in this thesis: the restriction to parametric covariate-treatment regressions and the missing comparison of the robustness of the outcome modeling approach and the assignment modeling approach to violations of the applied functional form assumption. Some of the limitations are incorporated in the suggested extensions of generalized analysis of covariance and the summarized research needs for the analysis of average total effects, described in the subsequent subsection.

5.4.1 Limitations of the Current Research

Shortcomings of the Monte Carlo Simulations In both parts of the simulation study we analyzed a fairly simple covariate-treatment regression with a single univariate covariate. Accordingly, the reported lower bounds for the required sample size necessary to obtain trustable standard errors for more complex covariate-

treatment regressions with multivariate and probably latent covariates need to be investigated and remain to be studied in detail.

Furthermore, the selected parameters for the data generation (see section 4.2) were not chosen in relation to typical effect sizes observed in specific research areas. The conduced statistical simulation showed, for instance, that ignoring the stochasticity of the covariates for outcome regressions with interaction terms yields biased standard errors for the *ATE*–estimator of the general linear model. The importance of this finding is of theoretical nature, admittedly the practical relevance cannot be judged based on the reported Monte Carlo results because we did not relate the parameters used for the data generation to realistic interaction effects typically observed for different content domains. Similarly, we must acknowledge that the appraisal of the practical importance of the reported failures of the (conventional) multi-group structural equation model (when the treatment group is considered as a fixed regressor) was beyond the scope of our simulation study.

Moreover, we did not investigate non-normality for data generation. The traditional analysis of covariance is known to be robust against violations of the normality assumption of the residuals with respect to the type-I-error rates (see, e. g., Rutherford, 2001, for a summary). Serious effects of non-normality for the traditional ANCOVA are expected only with respect to statistical power. It seems reasonable to assume that this result is generalizable to the regression estimate approach which is based on ordinary least-squares estimates of the covariate-treatment regression as well. Lei and Lomax (2005) found that the standard errors for structural equation models are not significantly affected by non-normality of the observed variables, but they reported that the accuracy of parameter estimates itself was sensitive to the distributions used in their simulation study. Raykov and Marcoulides (2006, p. 30) summarize the research to show that the maximum likelihood estimation for structural equation models can be applied for data with minor deviations from normality (see also Curran et al., 1996). With the help of a Monte Carlo simulation, the robustness of the maximum likelihood based implementations of generalized analysis of covariance developed in this thesis to skewed, heavy-tailed or in any other way non-normally distributed variables should be studied and compared, for instance, to the adjusted standard errors for the regression estimates.

Parametric Covariate-Treatment Regressions A major limitation of the considered implementations of generalized analysis of covariance is the focus on the simple linear parameterized intercept and effect function. Although a variety of approaches for the estimation of non-parametric regressions exist in the literature, which could easily be applied to the estimation of the difference in adjusted means, only linear effect functions were considered for the implementation of generalized analysis of covariance as structural equation models with nonlinear constraints. Non-parametric alternatives to the traditional analysis of covariance were mentioned in subsection 2.2.2. Promising approaches are discussed, for instance, an *ANCOVA*

based on cubic splines with fixed knots for the estimation of average total effects from nonlinear outcome regressions by R. J. A. Little et al. (2000) and Kang and Schafer (2007a). Nevertheless, especially for the suggested structural equation models, extensions to nonlinear intercept and effect functions need more research and are an important area for further investigations.

Robustness of the Outcome versus the Assignment Modeling Approach In general two modeling strategies can be applied for the estimation of (adjusted) average total effects. This thesis focused on the outcome modeling approach based on a linear parametrization of the covariate-treatment regression. According to Schafer and Kang (2008), an outcome regression model performs well in terms of bias, efficacy, and robustness when the prediction is strong. Therefore, an outcome regression with a covariate-treatment interaction was studied in this thesis under different levels of prediction strength. The results of the Monte Carlo study demonstrated that the proposed implementation of generalized analysis of covariance works in principle. Furthermore, we utilized the simulated datasets to illustrate the relative merits of the structural equation models compared to the general linear model. Nevertheless, the robustness of the covariate-treatment regression to misspecifications of the functional form assumption was not considered and especially not compared to alternative propensity score methods. Further research should extend the results given by Drake (1993), and might incorporate a discussion of tolerable misspecifications, i. e., valid transformations which do not invalidate the estimated average total effect (Waernbaum, 2010).

Further Limitations In this thesis, stochasticity of regressors and heterogeneity of residual variances were considered as challenges for the implementation of generalized analysis of covariance with covariate-treatment interactions. Of course a plethora of additional complications might arise when analyzing average total effects in real data applications. For instance, missing values, and especially missing values on the covariates, were not incorporated in the simulation study. Furthermore, the influence of outliers on the parameter estimates (see, e. g., Jureckova & Picek, 2006, for a derivation of the effect of outliers on the performance of estimators) should be studied because the outcome approach is expected to be notably sensitive to outliers when the functional form assumption of the covariate-treatment regression leads to heavy extrapolation due to limited covariate overlap (Imbens, 2004).

Moreover, the discussion of implications of the theory of stochastic causality for the design of observational studies was not in the scope of this thesis. A reasonable design of an observational study, for instance, non-zero treatment probabilities for each unit in each treatment condition, is probably the most important prerequisite for a successful estimation of average total effects from quasi-experimental designs. All adjustment procedures, i. e., propensity score based adjustment methods as well as regression based adjustment

methods, rest on similar causality conditions (see section 1.1.6). General advice for the design of observational studies can be found in, e. g., Steyer et al. (in press), Rubin (2008b), and Rosenbaum (2010).

Finally, a comparison of the asymptotic arguments applied by Schafer and Kang (2008) for the derivation of the variance approximation to the assumption used for deriving the approximated multi-group approach is still missing (see also J. Robins, Rotnitzky, & Zhao, 1994). With our simulation study we provided empirical evidence that the critical asymptotic covariances are essentially zero. It would be highly desirable to derive precise predictions about these asymptotic covariances from the underlying assumptions of the maximum likelihood estimation. Until now, the missing analytical justification of the augmentation approach poses a limitation for the absolute recommendation of the approximated multi-group model.

5.4.2 Further Extensions and Subsequent Research Questions

The different implementations of generalized analysis of covariance which have been developed as structural equation models with nonlinear constraints are of general interest for a variety of applications and subsequent research questions. As described in chapter 1, we focused mainly on the estimation of average total effects in quasi-experimental designs, i. e., on conditions where the fixed-X assumption is clearly inappropriate. The term *quasi-experimental designs* was used in keeping with previous publications by Steyer and Partchev (2008). In fact, the considered models apply very generally to *non-equivalent group designs* (Reichardt, 2005), to *quasi-experimental designs* (Shadish & Luellen, 2005) and accordingly to the general class of *observational studies* (Rosenbaum, 2005, 2010), as long as the empirical phenomenon can be described with the single-unit trial of a simple experiment or quasi-experiment. Even though no real data were empirically analyzed in this thesis, the studied generalized analysis of covariance can be recommended, for example, to the wider research field of *Program Evaluation* (see, for example, Mark, 2003).

Additionally, the core part of the discussed implementation of generalized analysis of covariance as structural equation model, i. e., nonlinear constraints of model parameters and Wald–tests to test (multiple) hypotheses about the estimated effects, can be utilized for statistical inference about further causal effects, for instance, various conditional total effects. In the following paragraphs we will suggest further extensions to broaden the practicability of generalized analysis of covariance as well as different research needs to develop the data analysis technique for the analysis of causal effects further.

Demonstration of the Effect of Measurement Error A brief discussion was given about the effect of covariates' measurement error on the estimated regression coefficients of the covariate-treatment regression in subsection 2.5.5. Furthermore, the considered structural equation models can be described as the generalization of the analysis of covariance to latent variables as suggested by Sörbom (1976, 1978), which has already been used in a number of studies (see, for example, Magidson, 1977, Bentler & Woodward, 1978,

Magidson & Sörbom, 1982, to name at least some of the earliest publications). Thus, the structural equation models with nonlinear constraints developed in this thesis are not only valuable for stochastic regressors and covariate-treatment interactions (benefit from the assumed joint distribution) but can be also easily applied when covariates are measured with error. Although the sample size requirements reported in section 4.8 might become lower limits due to the increased number of estimated parameters necessary for the additional measurement models, the developed structural equation models, if appropriately identified, are expected to give unbiased estimates of the (latent) regression coefficients and intercepts of the measurement model and should result in both unbiased average total effect estimators and correct standard errors for the estimated average total effects.

The underlying theory for generalized analysis of covariance with latent covariates is described in more detail by Steyer et al. (in press). In order to derive recommendations for applied research it would be beneficial to demonstrate the importance of latent covariates in the covariate-treatment regression, for example, by comparing generalized analysis of covariance with latent covariates as implemented in *EffectLite* (Steyer & Partchev, 2008) to the regression estimate approach (using only manifest indicator variables for the covariates). This would be of special interest for covariate-treatment regressions with interaction terms because the estimated regression coefficients for nonlinear terms (i. e., interaction terms, quadratic terms of covariates) are expected to be even more biased due to measurement error than the regression coefficients for the covariates themselves (see, e. g., Moosbrugger et al., in press).

Incorporation of the Assignment Model In order to make generalized analysis of covariance more robust against misspecifications of the functional form of the covariate-treatment regression, it would be of interest to study the incorporation of the assignment model into the developed structural equation modeling approaches. As a possible starting point for the development of a doubly robust generalized analysis of covariance, we suggest the extended regression estimates as described by Schafer and Kang (2008) [see also the various alternative strategies for doubly robust estimation based on regression estimates presented therein and in Kang & Schafer, 2007a]. Without changing either the definition or the computation of the average total effect estimator as regression estimate based on the predicted scores [see Equation (2.7)], the authors suggested, for example, to estimate different treatment-group-specific regressions with a weighted least-squares estimator, using $1/\hat{\pi}$ as weights in the treatment group and $1/(1-\hat{\pi})$ as weights in the control group. Therefore, a challenging research question would be, for instance, to compare the available weighted maximum likelihood (WML) approaches for the developed elaborated single group model (Asparouhov, 2004, 2005), with the extended regression estimate method based on weighted least-squares.

As an alternative to the weighted maximum likelihood estimation and as a replacement for a probably unstable simple inverse-weighting of the outcome variable, it might be a promising strategy to combine a

latent generalized analysis of covariance based on the developed structural equation models with nonlinear constraints and a subclassification approach based on indicator variables generated from the estimated propensity scores [see Equation (2.18) in section 2.2 on page 38]. Accordingly, the elaborated single group model should be investigated in more detail, in particular generalized by incorporating additional indicator variables for the different propensity score strata. The same strategy could be applied for the elaborated (or approximated) multi-group model as well, although this model is harder to specify and the estimates might well be less stable or empirically not identified for small sample sizes. Finally, see Hoshino (2007, 2008) for an integration of propensity scores into latent variable models based on Markov chain Monte Carlo methods. An integration of Hoshinos work into the tradition of the analysis of covariance as used in this thesis has not been done to this day and is an open question worthy of further investigation.

Beyond Binary Treatments The average total effect was considered throughout this thesis for a comparison of two treatment groups. Nevertheless, the structure of the presented models can be transferred to models with more than two treatment groups without further complications (see, e. g., Nagengast, 2006, for a discussion of the approximated multi-group model for multiple treatment groups). In section 3.3.2 it was shown that the Wald–test of a single constraint equals the test based on the standard error for the studied two group comparison. The mentioned flexibility of the multivariate δ-method comes into play when more than two groups are involved. The theory of stochastic causality is not restricted to cases with binary treatment variables.

Furthermore, generalizations of the estimation of causal effects for dose-response relationships are suggested within the Rubin Causal Model (see, e. g., Imbens, 2000; Wang & Donnan, 2001; Foster, 2003; Rosenbaum, 2003; Flores, 2004; Wasserman, 2004; Zanutto, Lu, & Hornik, 2005; Leon & Hedeker, 2005). Hence, an interesting avenue of research would be a generalization of the techniques for data analysis and statistical inferences studied here in line with the more general approaches presented, for instance, by Steyer (1992) and Steyer et al. (in press).

Nonlinear Effect Functions As mentioned in the previous subsection, the focus on linear parameterized intercept and effect functions is still a major limitation of the discussed generalized analysis of covariance (although covariate-treatment interactions are considered). Accordingly, we suggest continuing the research on orthogonalization of the covariates (see Flory, 2004, for a previous attempt) as an extension of the structural equation models developed in this thesis (see also T. D. Little et al., 2006, for a summary of the merits of orthogonalizing terms for the analysis of structural equation models, and Klein & Muthén, 2007, for estimating nonlinear covariate-treatment regressions). Although we did not study *randomization* or *permutation* tests as they are not robust against heterogeneity of residual variance in general (see Hayes,

1996), a comparison of parametric models with nonlinear effect functions and randomization based test statistics would be illuminating.

Multilevel-Structure and Stable Unit Treatment Value Assumption In chapter 1, we briefly mentioned an important part of the Rubin Causal Model, the *stable unit treatment value assumption* (*SUTVA*). Basically, *SUTVA* can be separated into two components: The postulation that no versions of treatments exist (related to the validity of the treatment variable X) and the assumption of "no interference between units". No interference between units is often understood as the assumption that the treatment assignment for one unit does not affect the outcome of another unit. This is of course a strong assumption which might be questionable in empirical applications even under randomization (see, for example, Staines, 2007). Possible violations of *SUTVA* for multi-level designs are frequently discussed, for example, by Rubin (1990); Gitelman (2005); Hong and Raudenbush (2005); Sobel (2006); Hong and Raudenbush (2006, 2008) and Nagengast (2009). Furthermore, throughout this thesis we have assumed that the treatment is assigned to individual units (see the mentioned single-unit trial in section 1.1) and that the sample for the observational study is obtained by simple random sampling. For example, VanderWeele (2008, p. 1940) points out that *"appropriate multilevel modeling techniques must be used to estimate the variance of the* $E[Y_{ik}|X_k = x, Z_{ik} = z, V_k = v]$ *estimates."*[1] Multi-level modeling (also known as hierarchical modeling, Raudenbush & Bryk, 2002) is especially necessary to obtain valid standard errors for the estimated regression coefficients. Furthermore, VanderWeele (2008) notes that bootstrapping techniques must be applied to obtain a valid standard error for an average total effect. An extension of the presented structural equation models with nonlinear constraints (including additional interactions between individual level or cluster level covariates and the treatment variable) was recently developed by Nagengast (2009). He also provides a discussion of the consequences resulting from multi-level designs for the theory of stochastic causality and gives the theoretical background for defining unbiasedness of the multilevel-covariate-treatment regression $E(Y|X,Z,V)$ with additional cluster-level covariates V.

Standard Error Bias Correction for Ordinary Least-Squares Estimator In section 3.2 we approached the variance of the *ATE*–estimator and predicted the underestimated standard error when a) covariate-treatment interactions are present and when b) the covariate(s) and the treatment variable are stochastic regressors at the same time. On the one hand our derivation explains under which conditions the ordinary least-squares regression is not robust against violations of the fixed-X assumption. On the other hand, the development of a correction for the underestimated standard error was impossible based on this analysis

[1] Y_{ik} denotes the post-treatment outcome Y for the individual i in the cluster k, Z_{ik} is the value of the individual-level covariate for the individual i in the cluster k, X_k is the value of the cluster treatment variable for cluster k, and V_k is the value of the cluster-level covariate for the cluster k. The covariates Z_{ik} and V_k are possibly multivariate.

without additional distributional assumptions (like multivariate normality which was assumed later for the maximum likelihood estimation of the developed structural equation models). In light of the recently published adjusted standard errors for regression estimates based on ordinary least-squares estimated group-specific covariate-regressions developed by Schafer and Kang (2008), this result is clearly unsatisfying. The observed performance of the regression estimate approach, even when covariates are stochastic, might reason further research to close the theoretical gap between the literature on ordinary least-squares estimation as presented in section 3.1.3 and the theoretical background used for the derivation of the adjusted standard errors for the *ATE*–estimator based on regression estimates.

Substantial Interaction Effects and Expected Heterogeneity of Residual Variance The final suggestion for subsequent research is motivated less by statistical or methodological needs than by the question to which extent the discovered standard error biases of misspecified generalized analysis of covariance are of practical importance. We studied the performance of different implementations under well-grounded properties of quasi-experimental designs as implied by the theory of stochastic causality. In order to bridge the gap between the theoretical considerations and the conditions of our simulation study on the one hand and the applied research about treatment evaluation on the other hand, it would be helpful to study the expected amount of covariate-treatment interactions in typical quasi-experimental evaluated treatments, at least in some selected content domains, for instance, in psychological or educational research. With the same goal in mind, reasonable expectations of typical amounts of between-group residual variance heterogeneity should, for example, be derived from the literature by meta-analytic techniques.

5.5 Practical Recommendations

5.5.1 General Linear Hypothesis and Mean-Centering

The general linear model for testing hypotheses about average total effects, including the general linear hypothesis and the moderated regression model with mean-centering based on the estimated means of the covariates, is generally not suggested if interaction terms are present. The test statistics for hypotheses of no average total effect based on the ordinary least-squares estimated covariate-treatment regression as well as the estimated variance of the *ATE*–estimator are found to be unreliable. Furthermore, all strategies based on the general linear model without a standard error correction are susceptible to heterogeneity of residual variances for unequal group sizes.

The studied regression estimate approach as described by Schafer and Kang (2008) with the corresponding adjusted standard errors overcomes both limitations of the general linear model and is recommended as implementation of generalized analysis of covariance for manifest covariates and small sample

sizes, in particular because of a slightly higher statistical power. However, the method does not trump the structural equation models with nonlinear constraints developed in this thesis, at least under the studied conditions for medium and large sample sizes, either with respect to the type-I-error rate or with respect to the statistical power.

5.5.2 Structural Equation Models with Nonlinear Constraints

Accurately specified structural equation models with nonlinear constraints can result in a correct standard error for the average total effect estimator, i. e., provide valid statistical inference for adjusted average total effects for an outcome modeling approach based on the covariate-treatment regression. We do not recommend the simple implementation of the analysis of covariance as a structural equation model which was suggested years ago by Sörbom (1978), although this model is still included as a standard example in textbooks and manuals (see, for example, Arbuckle, 2006). Likewise, single group models (including our developed elaborated single group model) are not advisable without restrictions. Based on the conditions studied in the Monte Carlo simulation, we favor the approximated multi-group structural equation model (Nagengast, 2006) as well as our elaborated multi-group model.

Hence, with respect to the software package *EffectLite* (Steyer & Partchev, 2008) the results of this thesis support the recommendation to incorporate the elaborated multi-group model as it is theoretically reasonable and applicable without any additional assumptions about asymptotic covariances between parameter estimates. If *EffectLite* is not available, the developed elaborated multi-group model is generally recommended for the estimation and testing of average total effects. Furthermore, a positive development would be for *EffectLite* to offer the option to compare the inference based on the approximated multi-group model with the conclusions obtained from the adjusted standard error of the regression estimation approach (for manifest covariates). We do not recommend the simple multi-group model available as *EffectLite* option for observational studies or quasi-experimental designs because of missing robustness observed for stochastic treatment variables.

The suitability of generalized analysis of covariance as a structural equation model with nonlinear constraints for daily use depends on the adequacy of the assumption of Z-conditional unbiasedness of the covariate-treatment regression. The models presented in this thesis can be utilized for empirical applications and might serve as the starting point for various enhancements to the specific requirements of quasi-experimental studies, if Z-conditional unbiasedness can be assumed and if the covariate-treatment regression is correctly specified as structural equation model. However, if the complete set of covariates, that influences both the treatment assignment and the outcome variable, is not available or if the covariate-treatment regressions' functional form is in doubt, the following statement still seems valid. *"...evaluators*

may wish to apply various techniques to the data to determine whether the conclusions differ depending on the different analytic assumptions underlying the techniques. [...] we believe that multiple analyses are necessary in this case [i.e., for quasi-experimental data], no strategy is sufficient to assure that all relevant confounds have been appropriately taken into account. Ultimately, one must rely on theory to help interpret the results. The weaker the design of the study, the heavier the burden of interpretation must rest with theory." (Magidson & Sörbom, 1982, p. 321) The developed structural equation models with nonlinear constraints may add to the repertoire of the available adjustment methods for quasi-experimental designs.

Bibliography

Abadie, A., & Imbens, G. W. (2006). *On the failure of the bootstrap for matching estimators* (Tech. Rep.). John F. Kennedy School of Government, Harvard University, Retrieved March 1, 2008, from http://www.hks.harvard.edu/fs/aabadie/bootstrap.pdf.

Agresti, A. (2007). *An introduction to categorical data analysis.* Hoboken, NJ: John Wiley & Sons.

Aguinis, H., Petersen, S. A., & Pierce, S. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods*, *2*, 315–339.

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, *62*, 488–499.

Aiken, L. S., & West, S. G. (1996). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage Publications.

Akritas, M. G., Arnold, S. F., & Du, Y. (2000). Nonparametric models and methods for nonlinear analysis of covariance. *Biometrika*, *87*, 507–526.

Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin*, *115*, 308–314.

Algina, J., & Olejnik, S. (2000). Determining sample size for accurate estimation of the squared multiple correlation coefficient. *Multivariate Behavioral Research*, *35*, 119–137.

Allison, P. D. (1995). The impact of random predictors on comparisons of coefficients between models: Comment on Clogg, Petkova, and Haritou. *The American Journal of Sociology*, *100*, 1294–1305.

Altman, D. G. (1998). Adjustment for covariate imbalance. In P. Armitage & T. Colton (Eds.), *Encyclopaedia of biostatistics* (pp. 1000–1005). New York, NY: John Wiley & Sons.

Angrist, J. D., & Krueger, A. B. (1999). Empirical strategies in labor economics. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3A). Amsterdam: Elsevier.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Presss.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum.

Arbuckle, J. L. (2006). *Amos 7.0 user's guide*. Chicago, IL: SPSS.

Asparouhov, T. (2004). *Weighting for unequal probability of selection in multilevel modeling* (Tech. Rep.). Mplus Web Notes: No. 8, Retrieved from http://www.statmodel.com/examples/webnote.shtml.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411–434.

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

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statistics in Medicine*, *25*, 2084–2106.

Ayinde, K. (2007). A comparative study of the performances of the OLS and some GLS estimators when stochastic regressors are both collinear and correlated with error terms. *Journal of Mathematics and Statistics*, *3*, 196–200.

Bagozzi, R. P., & Yi, Y. (1989). On the use of structural equation models in experimental designs. *Journal of Marketing Research*, *26*, 271–284.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–972.

Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Helath*, *9*, 377–385.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal*, *2*, 358–377.

Belsley, D. A. (1984). Demeaning conditioning diagnostics through centering. *The American Statistician*, *38*, 73–77.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. Hoboken, NJ: John Wiley & Sons.

Benjamin, D. J. (2003). Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. *Journal of Public Economics*, *87*, 1259–1290.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, *31*, 419–456.

Bentler, P. M. (1991). Modeling of intervention effects. In C. G. Leukefeld & W. J. Bukoski (Eds.), *Drug abuse prevention intervention research: Methodological issues.* Rockville, MD: National Institute on Drug Abuse.

Bentler, P. M. (1995). EQS structural equations program manual [Computer software manual]. Encino, CA: Multivariate Software.

Bentler, P. M., & Woodward, J. A. (1978). A head start reevaluation: Positive effects are not yet demonstrable. *Evaluation Review, 2*, 493–510.

Berry, W. D. (1993). *Understanding regression assumptions.* Thousand Oaks, CA: Sage Publications.

Bini, M., Monari, P., Piccolo, D., & Salmaso, L. (Eds.). (2010). *Statistical methods for the evaluation of educational services and quality of products.* Berlin: Physica.

Bishop, Y. M. M., Fienberg, E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice.* Cambridge, MA: MIT Press.

Blundell, R., & Costa Dias, M. (2000). Evaluation methods for non-experimental data. *Fiscal Studies, 21*, 427–468.

Bobko, P., & Russell, C. (1990). Variance homogeneity in interactive regression: A clarifying note about data transformations. *Journal of Applied Psychology, 75*, 95–96.

Bollen, K. A. (1989). *Structural equation modeling with latent variables.* New York, NY: John Wiley & Sons.

Bollen, K. A. (1996). A limited-information estimator for lisrel models with or without heteroscedastic errors. In G. Marcoulides & R. Schumaker (Eds.), *Advanced structural equation modling - issues and techniques* (pp. 227–241). Mahwah, NJ: Lawrence Erlbaum.

Bollen, K. A., & Paxton, P. (1998). Interactions of latent variables in structural equation models. *Structural Equation Modeling, 5*, 267–293.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In *Structural Equation Modeling: Present and Future. A Festschrift in honor of Karl Jöreskog* (pp. 139–168). Chicago, CA: Scientific Software International.

Bowman, A. W., & Azzalini, A. (2007). R package `sm`: nonparametric smoothing methods (version 2.2) [Computer software manual]. University of Glasgow, UK and Università di Padova, Italia. Available from `http://www.stats.gla.ac.uk/~adrian/sm`

Brand, J. E., & Halaby, C. N. (2006). Regression and matching estimates of the effects of elite college attendance on educational and career achievement. *Social Science Research, 35*, 749–770.

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance.* Newbury Park, CA: Sage Publications.

Bryck, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*, 396–404.

Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in ols regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics, 33*, 21–40.

Caliendo, M. (2006). *Microeconometric evaluation of labour market policies*. Berlin: Springer.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally.

Cantoni, E., & De Luna, X. (2006). Non-parametric adjustment for covariates when estimating a treatment effect. *Nonparametric Statistics, 18*, 227–244.

Carpenter, J., Kenward, M., & Vansteelandt, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *Journal of the Royal Statistical Society Series A, 169*, 571–584.

Carroll, R. J., & Ruppert, D. (1988). *Transformation and weighting in regression*. New York, NY: Chapman & Hall.

Cepeda, M., Boston, R., Farrar, J., & Strom, B. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology, 158*, 280–287.

Chambers, J. M., & Hastie, T. J. (Eds.). (1991). *Statistical models in S*. London: Chapman & Hall.

Chen, X. (2006). The adjustment of random baseline measurements in treatment effect estimation. *Journal of Statistical Planning and Inference, 136*, 4161–4175.

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health Nations Health, 43*, 684–91.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 13*, 261–281.

Cochran, W. G. (1968a). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*, 295–313.

Cochran, W. G. (1968b). Errors of measurement in statistics. *Technometrics, 10*, 637–666.

Cochran, W. G. (1977). *Sampling techniques*. New York, NY: John Wiley & Sons.

Cochran, W. G. (1983). *Planning and analysis of observational studies*. New York, NY: John Wiley & Sons.

Cochran, W. G., & Chambers, S. (1965). The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A, 128*, 234–255.

Cohen, J. (1978). Partialed products are interactions; Partialed powers are curve components. *Psychological Bulletin, 85*, 858–866.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*, 174–184.

Cook, D. R., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, *70*, 1–10.

Cook, T. D., & Campbell, D. C. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*, 724–750.

Cook, T. D., Steiner, P. M., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability, and mode of data analysis: Results from two types of within-study comparisons. *Multivariate Behavioral Research*, *44*, 828–847.

Cox, D. R., & McCullagh, P. (1982). A biometrics invited paper with discussion. Some aspects of analysis of covariance. *Biometrics*, *38*, 541–561.

Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics*, *43*, 895–901.

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, *45*, 215–233.

Cribbie, R. A., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*, *60*, 893–907.

Cribbie, R. A., & Jamieson, J. (2004). Decreases in posttest variances and the measurement of change. *Methods of Psychological Research Online*, *9*, 37–55.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

D'Agostino, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281.

D'Agostino, R. B., Lang, W., Walkup, M., Morgan, T., & Karter, A. (2001). Examining the impact of missing data on propensity score estimation in determining the effectiveness of self-monitoring of blood glucose (SMBG). *Health Services & Outcomes Research Methodology*, *3–4*, 291–315.

D'Agostino, R. B., & Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, *95*, 749–759.

Darken, P. F. (2004). Evaluating assumptions for least squares analysis using the general linear model: A guide for the pharmaceutical industry statistician. *Journal of Biopharmaceutical Statistics in Medicine*, *14*, 803–816.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. New York, NY: Springer.

Davey, A., & Savla, J. (2010). *Statistical power analysis with missing data: A structural equation modeling approach*. New York, NY: Routledge.

Degracie, J. S., & Fuller, W. A. (1972). Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association*, *67*, 930-937.

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

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

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

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1236.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York, NY: John Wiley & Sons.

Enders, C. K. (2001a). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, *61*, 713–740.

Enders, C. K. (2001b). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling*, *8*, 128–141.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457.

Engqvist, L. (2005). The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Animal Behaviour*, *70*, 967–971.

Eye, A. von, & Schuster, C. (1998). *Regression analysis for social sciences*. San Diego, CA: Academic Press.

Farley, J. U., & Reddy, S. K. (1987). A factorial evaluation of effects of model specification and error on parameter estimation in a structural equation model. *Multivariate Behavioral Research*, *22*, 71–90.

Feldt, L. S. (1958). A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, *23*, 335–353.

Finney, J. W., Mitchell, R. E., Cronkite, R. C., & Moos, R. H. (1984). Methodological issues in estimating main and interactive effects: Examples from coping/social support and stress field. *Journal of Health and*

Social Behavior, *25*, 85–98.

Fisher, R. A. (1925/1954). *Statistical methods for research workers* (12th ed.). New York, NY: Hafner Publishing.

Fisicaro, S. A., & Tisak, J. (1994). A theoretical note on the stochastics of moderated multiple regression. *Educational and Psychological Measurement*, *54*, 32–41.

Flores, C. A. (2004). *Estimation of dose-response functions and optimal doses with a continuous treatment* (Working Papers). University of Miami, Department of Economics. Available from `http://ideas.repec.org/p/mia/wpaper/0707.html`

Flory, F. (2004). *Estimating and testing average causal effects in regression models with interactions and stochastic predictor variables.* Unpublished Diploma thesis, University of Jena, Jena, Thuringia, Germany.

Flory, F. (2008). *Average treatment effects in structural equation models with interactions between treatment and manifest or latent covariates.* Unpublished doctoral dissertation, University of Jena, Jena, Thuringia, Germany.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care*, *41*, 1183–1192.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation Review*, *32*, 392–409.

Froelich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *The Review of Economics and Statistics*, *86*, 77–90.

Fuller, W. A. (1987). *Measurement error models.* New York, NY: John Wiley & Sons.

Gallin, J. (1983). Misspecifications that can result in path analysis structures. *Applied Psychological Measurement*, *7*, 125–137.

Ganju, J. (2004). Some unexamined aspects of analysis of covariance in pretest-posttest studies. *Biometrics*, *60*, 829–833.

Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, *106*, 516–524.

Gelman, A. (2004). Treatment effects in before-after data. In A. Gelman & X.-L. Meng (Eds.), *Applied bayesian modeling and causal inference from incomplete-data perspectives.* (pp. 195–202). Chichester: John Wiley & Sons.

Gelman, A., & Hill, J. L. (2007). *Data analysis using regression and multilevel / hierarchical models.* New York: New York, NY: Cambridge University Press.

Gelman, A., & Meng, X.-L. (2004). *Applied bayesian modeling and causal inference from incomplete-data perspectives*. Chichester: John Wiley & Sons.

Gelman, A., & Pardoe, I. (2007). Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, *37*, 23–51.

Gitelman, A. I. (2005). Estimating causal effects from multilevel group-allocation data. *Journal of Educational and Behavioral Statistics*, *30*, 397–412.

Glazerman, S., Levy, D., & Myers, D. (2003). Nonexperimental versus experimental estimates of earning impacts. *The Annals of the American Academy*, *589*, 63–93.

Glueck, D. H., & Muller, K. E. (2003). Adjusting power for a baseline covariate in linear models. *Statistics in Medicine*, *22*, 2535–2551.

Glynn, A. N., & Quinn, K. M. (2009). Causalgam: Estimation of causal effects with generalized additive models [Computer software manual]. Available from `http://CRAN.R-project.org/package=CausalGAM` (R package version 0.1-2)

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*, 36–56.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling*, *10*, 80–100.

Greene, W. (2007). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances and algorithms. *Journal of Computational and Graphical Statistics*, *2*, 405–420.

Guo, S., & Fraser, M. W. (2010). *Propensity score analysis*. Thousand Oaks, CA: Sage Publications.

Hancock, G. R., & Mueller, R. O. (2006). *Structural equation modeling : A second course*. Greenwich, CT: Information Age Publishing.

Hartenstein, S. (2005). *ANOVA and MANOVA using likelihood ratio tests. A simulation study*. Unpublished Diploma thesis, University of Jena, Jena, Thuringia, Germany.

Harwell, M. (2003). Summarizing monte carlo results in methodological research: The single-factor, fixed-effects ANCOVA case. *Journal of Educational and Behavioral Statistics*, *28*, 45–70.

Harwell, M., & Serlin, R. C. (1988). An empirical study of a proposed test of nonparametric analysis of covariance. *Psychological Bulletin*, *104*, 268–281.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman & Hall.

Haviland, A., Nagin, D. S., & Rosenbaum, P. R. (2007). Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods*, *12*, 247–267.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing $H_0 : p = 0$. *Psychological Methods*, *1*, 184–198.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709–722.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions – An overview. *Journal of Econometrics*, *30*, 239–267.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In J. J. Heckman & E. E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6B). Amsterdam, NL: Elsevier.

Henderson, C. R. (1982). Analysis of covariance in the mixed model: Higher-level, nonhomogeneous, and random regressions. *Biometrics*, *38*, 623–640.

Hernán, M. A., Hernández-Díaz, S., Werler, M. M., & Mitchell, A. A. (2002). Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, *155*, 176–184.

Hill, J. (2008). Comment: Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *484*, 1346–1350.

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment efects using propensity score matching. *Statistics in Medicine*, *25*, 2230-2256.

Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, *3–4*, 259–278.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2005). Matchit: Nonparametric preprocessing for parametric causal inference [Computer software manual]. Retrieved from http://GKing.Harvard.Edu/MatchIt.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, *27*, 205–224.

Hong, G., & Raudenbush, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inferrence for multilevel observational data. *Journal of the American Statistical Association*, *101*, 901–910.

Hong, G., & Raudenbush, S. W. (2008). Causal inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, *33*, 333–362.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685.

Hoshino, T. (2007). Doubly robust-type estimation for covariate adjustment in latent variable modeling. *Psychometrika*, *72*, 535–549.

Hoshino, T. (2008). A bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, *52*, 1413–1429.

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge, UK: Cambridge University Press.

Härdle, W., & Linton, O. (1994). Applied nonparametric methods. In R. F. Engle & D. McFadden (Eds.), *Handbook of econometrics* (Vol. 4). Amsterdam, NL: North-Holland.

Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherpy outcome research:. *Journal of Consulting and Clinical Psychology*, *57*, 131–137.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York, NY: John Wiley & Sons.

Hunka, S., & Leighton, J. (1997). Defining Johnson-Neyman regions of significance in the three-covariate ANCOVA using mathematica. *Journal of Educational and Behavioral Statistics*, *22*, 361–387.

Huppler Hullsiek, K., & Louis, T. A. (2002). Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics*, *2*, 179–193.

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society A*, *171*, 481–502.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*, 706–710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*, 5–86.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. In P. R. Krishnaiah (Ed.), *Handbook of statistics* (Vol. 1, pp. 199–236). Amsterdam, NL: North-Holland.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications.

Jamieson, J. (2004). Analysis of covariance ANCOVA with difference scores. *International Journal of Psychophysiology*, *52*, 277–283.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202.

Jöreskog, K. G., & Sörbom, D. (1988). LISREL 7: A Guide to the Program and Applications [Computer software manual]. Lincolnwood, IL: Scientific Software International: Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996 - 2001). LISREL 8: User's Reference Guide (2nd ed.) [Computer software manual]. Lincolnwood, IL: Scientific Software International.

Jöreskog, K. G., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. Marcoulides & R. Schumaker (Eds.), *Advanced structural equation modling – Issues and techniques* (pp. 57–81). Mahwah, NJ: Lawrence Erlbaum.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115–134.

Jureckova, J., & Picek, J. (2006). *Robust statistical methods with R.* Boca Raton, FL: Chapman & Hall.

Kang, J. D. Y., & Schafer, J. L. (2007a). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean for incomplete data. *Statistical Science*, *22*, 523–539.

Kang, J. D. Y., & Schafer, J. L. (2007b). Rejoinder: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 574–580.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research*, *23*, 69–86.

Kaplan, D. (1989). A study of the sampling variability of the z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*, 41–57.

Kaplan, D. (2000). *Structural equation modeling – Foundations and extensions.* Thousand Oaks, CA: Sage Publications.

Karpman, M. B. (1983). The Johnson-Neyman technique using SPSS or BMDP. *Educational and Psychological Measurement*, *43*, 137–147.

Keele, L. (2008). *Semiparametric regression for the social sciences.* Chichester: John Wiley & Sons.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201–210.

King, G., & Zeng, L. (2001). Improving forecasts of state failure. *World Politics*, *53*, 623–58.

King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, *14*, 131–159.

Kish, L. (1965). *Survey sampling.* New York, NY: John Wiley & Sons.

Kleiber, C., & Zeileis, A. (2008). *Applied econometrics with R.* New York: New York, NY: Springer.

Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*, 457–474.

Klein, A. G., & Muthén, B. O. (in press). Quasi maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*.

Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*, 647–673.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Kotz, S., Read, C. B., Balakrishnan, N., & Vidakovic, B. (Eds.). (2005). *Encyclopedia of statistical sciences*. New York, NY: John Wiley & Sons.

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, *13*, 141–151.

Krauth, J. (2000). *Experimental design: A handbook and dictionary for medical and behavioral research*. Amsterdam, NL: Elsevier.

Kröhne, U., & Wolf, A. (2002). *Ein struktureller Ansatz zur Erklärung von Unterschieden zwischen computerisierten und Papier-und-Bleistift Tests [A structured approach toward the explanation of differences between computerized and paper-pencil tests]*. Unpublished Diploma thesis, University of Jena, Jena, Thuringia, Germany.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement*, *58*, 42–67.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., et al. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, *163*, 262–270.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical model* (5th ed.). New York, NY: McGraw-Hill.

Larholt, K. M., & Sampson, A. R. (1995). Effects of heteroscedasticity upon certain analyses when the regression lines are not parallel. *Biometrics*, *51*, 731–737.

Lee, B. K., Lessler, J., & Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine*, doi: 10.1002/sim.3782.

Lee, S.-Y. (2007). *Structural Equation Modeling - A Bayesian Approach*. Chichester: John Wiley & Sons.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*, 1–27.

Leon, A. C., & Hedeker, D. (2005). A mixed-effects quintile-stratified propensity adjustment foreffectiveness analyses of ordered categorical doses. *Statistics in Medicine*, *24*, 647–658.

Leuven, E., & Sianesi, B. (2003). *PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.* Retrieved from http://ideas.repec.org/r/boc/bocode/s432001.html.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Lin, D., Psaty, B., & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, *54*, 948–963.

Little, R. J., Long, Q., & Lin, X. (2008). Comment: Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *484*, 1344-1346.

Little, R. J. A., An, H., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods*, *5*, 459–476.

Little, R. J. A., & An, H. G. (2004). Robust likelihood-based analysis of multivariate data with missingvalues. *Statistica Sinica*, *14*, 949–968.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, *13*, 497-519.

Lohr, S. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury Press.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224.

Long, J. S., & Trivedi, P. K. (1992). Some specification tests for the linear regression model. *Sociological Methods Research*, *21*, 161–204.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 307–321.

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review*, *29*, 530–558.

Lunceford, J., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, *23*, 2937–2960.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Mahwah, NJ: Lawrence Erlbaum.

MacKinnon, D. P., Lockwood, C. P., Hoffmann, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.

MacKinnon, J. G., & White, H. (1985). Some heteroskedastic-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*, 305–325.

Maddala, G. S. (1992). *Introduction to econometrics*. New York, NY: Macmillan.

Magidson, J. (1977). Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: A general alternative to ANCOVA. *Evaluation Review, 1*, 399–420.

Magidson, J., & Sörbom, D. (1982). Adjusting for confounding factors in quasi-experiments: Another reanalysis of the westinghouse head start evaluation. *Educational Evaluation and Policy Analysis, 4*, 321–329.

Mark, M. M. (2003). Program evaluation. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology: Research methods in psychology* (Vol. 2). Hoboken, NJ: John Wiley & Sons.

Marquardt, D. W. (1980). A critique of some ridge regression methods: Comment. *Journal of the American Statistical Association, 75*, 87–91.

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*, 275–300.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Maxwell, S. E., Delaney, H. D., & O'Callaghan, M. F. (1993). Analysis of covariance. In K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (Vol. 137, pp. 63–104). New York, NY: Marcel Dekker.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods, 9*, 403–425.

McKinlay, S. M. (1975). The design and analysis of the observational study – A review. *Journal of the American Statistical Association, 70*, 503–520.

Mendoza, J. L., & Stafford, K. L. (2001). Correlation coefficient under the fixed and random regression models: A computer program and confidence intervals, power calculation, and sample size estimation for the squared multiple useful standard tables. *Educational and Psychological Measurement, 61*, 650–667.

Milliken, G. A., & Johnson, D. E. (2002). *Analysis of messy data, volume III: Analysis of covariance*. London: Chapman & Hall.

Miyazaki, Y., & Maier, K. S. (2005). Johnson-Neyman type technique in hierarchical linear models. *Journal of Educational and Behavioral Statistics, 30*, 233–259.

Moore, K. L., & Laan, M. J. van der. (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, *28*, 39–64.

Moosbrugger, H., Schermelleh-Engel, K., A., K., & Klein, A. G. (in press). Testing multiple nonlinear effects in structural equation modeling: A comparison of alternative estimation approaches. In T. Teo & M. S. Khine (Eds.), *Structural equation modelling in educational research: Concepts and applications*. (chap. 6). Rotterdam, NL: Sense Publishers.

Morgan, S. L., & Harding, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research*, *35*, 3–60.

Morgan, S. L., & Todd, J. J. (2008). A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, *38*, 231–281.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference – Methods and principles for social research*. New York: New York, NY: Cambridge University Press.

Mulaik, S. A. (2009). *Linear causal modeling with structural equations*. Boca Raton, FL: Chapman & Hall.

Muller, K. E., & Stewart, P. W. (2006). *Linear model theory: Univariate, multivariate, and mixed models*. Hoboken, NJ: John Wiley & Sons.

Muthén, B. (1998-2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81–117.

Muthén, B. O., & Asparouhov, T. (2002). *Modeling of heteroscedastic measurement errors* (Tech. Rep.). Mplus Web Notes: No. 3, Retrieved from http://www.statmodel.com/examples/webnote.shtml.

Muthén, B. O., & Asparouhov, T. (2003). *Modeling interactions between latent and observed continuous variables using maximum-likelihood estimation in mplus* (Tech. Rep.). Mplus Web Notes: No. 6, Retrieved from http://www.statmodel.com/examples/webnote.shtml.

Muthén, L. K., & Muthén, B. O. (1998 - 2007). Mplus User's Guide (5th ed.) [Computer software manual]. Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599–620.

Myers, J. L., & Well, A. D. (2003). *Research design and statistical analysis*. Mahwah, New Jersey: Mahwah, NJ: Lawrence Erlbaum.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.

Nagengast, B. (2006). *Standard errors of ACE estimates: Comparing adjusted group means against the adjusted grand mean. A simulation study*. Unpublished diploma thesis, University of Jena, Jena, Thuringia, Germany.

Nagengast, B. (2009). *Causal inference in multilevel models*. Unpublished doctoral dissertation, University of Jena, Jena, Thuringia, Germany.

Neter, J., Wasserman, W., & Kutner, M. H. (1983). *Applied linear regression models*. Homewood, IL: Richard D. Irwin.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*, 439 – 478.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science, 5*, 465–480.

O'Connor, B. P. (1988). All-in-one programs for exploring interactions in moderated multiple regression. *Educational and Psychological Measurement, 8*, 833–837.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician, 46*, 27–29.

Olsson, U. H., Foss, T., Troye, S. V., & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling, 7*, 557–595.

Parsons, L. S. (2004). Performing a 1:n case-control match on propensity score. In *SUGI 29 retrieved november 26, 2008, from http://www2.sas.com/proceedings/sugi29/toc.html*.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press.

Pearl, J. (2003). Statistics and causal inference: A review. *TEST, 12*, 281–345.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys, 3*, 96–146.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (2009). Unbiased causal inference from an observational study: Results of a within-study comparison. *Educational Evaluation and Policy Analysis, 31*, 463–479.

Pohl, S., Steiner, P. M., Eisermann, J., Soellner, R., & Cook, T. D. (submitted). Empirically demonstrated unbiased causal inference from an observational study: A replication of the within-study comparison of Shadish, Clark and Steiner.

Posner, M. A., Ash, A. S., Freund, K. M., Moskowitz, M. A., & Shwartz, M. (2001). Comparing standard regression, propensity score matching, and instrumental variables methods for determining the influence of mammography on stage of diagnosis. *Health Services & Outcomes Research Methodology, 2*, 279–290.

Potthoff, R. F. (1964). On the Johnson-Neyman technique and some extensions thereof. *Psychometrika, 29*, 241–256.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and*

Behavioral Statistics, *31*, 437–448.

Qi, L., Little, R. J. ., & Lin, X. (2008). Causal inference in hybrid intervention trials involving treatment choice. *Journal of the American Statistical Association*, *103*, 474–484.

Quinn, G., & Keough, M. (Eds.). (2002). *Experimental design and data analysis for biologists*. New York, NY: Cambridge University Press.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B*, *31*, 350–371.

Rao, C. R. (1973). *Linear statistical inference and its applications*. London: New York, NY: John Wiley & Sons.

Rao, C. R., & Toutenburg, H. (1999). *Linear models: Least squares and alternatives*. New York, NY: Springer.

Raudenbush, S. W., & Bryk, A. S. (Eds.). (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Raykov, T., & Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Structural Equation Modeling*, *11*, 621-637.

Raykov, T., & Marcoulides, G. A. (2006). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum.

Reichardt, C. S. (2005). Nonequivalent group design. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Chichester: John Wiley & Sons.

Rencher, A. C., & Schaalje, G. B. (2007). *Linear model in statistics* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Ridgeway, G., McCaffrey, D., & Morral, A. (2006). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the `twang` package [Computer software manual].

Ridgeway, G., & McCaffrey, D. F. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 540–543.

Robins, J., & Greenland, S. (1986). The role of model selection in causal inference from non-experimental data. *American Journal of Epidemiology*, *123*, 392–402.

Robins, J., Rotnitzky, A., & Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.

Robins, J. M., Mark, S. D., & Newey, W. K. (1992). Estimating exposure effects by modelling the expectations of exposure coditional on confounders. *Biometrics*, *48*, 479–495.

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression-models with missing data. *Journal of the American Statistical Association*, *90*, 122–129.

Robins, J. M., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, *22*, 544–559.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*, 931–954.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within group regressions. *Educational and Psychological Measurement*, *41*, 73–84.

Rosenbaum, P. R. (1986). Dropping out of high-school in the united-states – An observational study. *Journal of Educational Statistics*, *11*, 207–224.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*, 387–394.

Rosenbaum, P. R. (1998). Propensity score. In P. Armitage & E. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 4267–4272). New York, NY: John Wiley & Sons.

Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, *97*, 183–192.

Rosenbaum, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, *17*, 286–304.

Rosenbaum, P. R. (2002c). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R. (2003). Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics*, *4*, 1–10.

Rosenbaum, P. R. (2005). Observational study. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1451–1462). Chichester: John Wiley & Sons.

Rosenbaum, P. R. (2010). *Design of observational studies*. New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, *45*, 212–218.

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

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

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control-group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, *39*, 33–38.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation : A systematic approach*. Thousand Oaks, CA: Sage Publications.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, *29*, 159–183.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–590.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1–26.

Rubin, D. B. (1978). Bayesian-inference for causal effects: The role of randomization. *Annals of Statistics*, *6*, 34–58.

Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, *36*, 293–298.

Rubin, D. B. (1986). Statistics and causal inference - Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961–962.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons.

Rubin, D. B. (1990). Formal modes of statistical-inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279–292.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *3–4*, 169–188.

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

Rubin, D. B. (2008a). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, *484*, 1350–1353.

Rubin, D. B. (2008b). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*, 808–840.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, *52*, 249–264.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*, 573–585.

Rücker, G. (1989). A two-stage trial design for testing treatment, self-selection, and treatment preference effects. *Statistics in Medicine*, *8*, 477–485.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge, UK: Cambridge University Press.

Rutherford, A. (2001). *Introducing ANOVA and ANCOVA: A GLM approach*. London: Sage Publications.

Ryan, T. P. (1996). *Modern regression methods*. New York, NY: John Wiley & Sons.

Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association, 69*, 682–689.

Sarkar, D. (2008). *Lattice – Multivariate data visualization with R*. New York, NY: Springer.

Satorra, A., & Saris, W. E. (1985, March). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50*, 83–90.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.

Schafer, J. L., & Kang, J. D. (2008). Average causal effects from observational studies: A practical guide and simulated example. *Psychological Methods, 13*, 279–313.

Scheffe, H. (1959). *The analysis of variance*. New York, NY: John Wiley & Sons.

Scheiner, S. M., & Gurevitch, J. (Eds.). (2001). *Design and analysis of ecological experiments*. New York, NY: Oxford University Press.

Schochet, P. Z. (2009). Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference, 140*, 246–259.

Schumacker, R. E. (2002). Latent variable interaction modeling. *Structural Equation Modeling, 9*, 40–54.

Schumacker, R. E., & Marcoulides, G. A. (Eds.). (1998). *Interaction and non-linear effects in structural equation*. Mahwah, NJ: Lawrence Erlbaum.

Sekhon, J. S. (2009). Opiates for the matches: Matching methods for causal inference. *Annual Review of Political Science, 12*, 487–508.

Senn, S., Graf, E., & Caputo, A. (2007). Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. *Statistics in Medicine, 26*, 5529–5544.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008a). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 103*, 1334–1344.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008b). Rejoinder: Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association, 484*, 1353–1356.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology, 60*, 607–629.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shadish, W. R., & Luellen, J. K. (2005). Quasi-experimental designs. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1641–1644). Chichester: John Wiley & Sons.

Shaha, B. R., Laupacisa, A., Huxa, J. E., & Austina, P. C. (2005). Propensity score methods gave similar results to traditional regression modeling in observational studies: A systematic review. *Journal of Clinical Epidemiology*, *58*, 550–559.

Shaikh, A. M., Simonsen, M., Vytlacil, E. J., & Yildiz, N. (2009). A specification test for the propensity score using its distribution conditional on participation. *Journal of Econometrics*, *151*, 33–46.

Sheather, S. J. (2009). *A modern approach to regression with R*. New York, NY: Springer.

Shieh, G. (2003). A comparative study of power and sample size calculations for multivariate general linear models. *Multivariate Behavioral Research*, *38*, 285-307.

Shieh, G. (2006). Exact interval estimation, power calculation, and sample size determination in normal correlational analysis. *Psychometrika*, doi: 10.1007/s11336-004-1221-6.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. Clogg C., & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1–38). New York, NY: Plenum Press.

Sobel, M. E. (1998). Causal inference in statistical models of the process of socioeconomic achievement: A case study. *Sociological Methods & Research*, *27*, 318–348.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *Journal of the American Statistical Association*, *101*, 1398–1407.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

Sörbom, D. (1976). Advances in psychological and educational measurement. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), (pp. 159–170). New York, NY: John Wiley & Sons.

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika*, *43*, 381–396.

Staines, G. L. (2007). Comparative outcome evaluations of psychotherapies: Guidelines for addressing eight limitations of the gold standard of causal inference. *Psychotherapy: Theory, Research, Practice, Training*, *44*, 161–174.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle [Theory of causal regression models]*. Stuttgart: Fischer.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression [Probability and regression]*. Berlin: Springer.

Steyer, R., Fiege, C., & Rose, N. (2010). Analyzing total, direct and indirect causal effects in intervention studies.

Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbaisedness. *Methods of Psychological Research Online*, *5*, 55–87.

Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C., & Buhl, T. (2000). Causal regression models I: Individual and average causal effects. *Methods of Psychological Research Online*, *5*, 39–71.

Steyer, R., Nachtigall, C., Wüthrich-Martone, O., & Kraus, K. (2002). Causal regression models III: Covariates, conditional, and unconditional average causal effects. *Methods of Psychological Research Online*, *7*, 41–68.

Steyer, R., & Partchev, I. (2008). *EffectLite for Mplus: A program for the uni- and multivariate analysis of unconditional, conditional and average mean differences between groups [Computer software and manual]*. Retrieved May 5, 2008, from www.statlite.com.

Steyer, R., Partchev, I., Kröhne, U., Nagengast, B., & Fiege, C. (in press). *Probability and causality*. Heidelberg: Springer.

Steyer, R., Wolf, A., Funke, F., & Partchev, I. (2009). Strukturgleichungsmodelle [structural equation models]. In H. Holling & R. Schwarzer (Eds.), *Handbuch der Psychologie, Evaluationsforschung: Band 1: Modelle und Methoden*. Hogrefe.

Stürmer, T., Joshia, M., Glynn, R. J., Avorna, J., & Rothman, S., Kenneth J.and Schneeweiss. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*, *59*, 437–447.

Stuart, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of 'A critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Statistics in Medicine*, *27*, 2062–2065.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, *101*, 1619–1637.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, *22*, 560–568.

Tang, M.-L., & Bentler, P. M. (1998). Theory and method for constrained estimation in structural equation models with incomplete data. *Computational Statistics & Data Analysis*, *27*, 257–270.

Tate, R. (2004). Interpreting hierarchical linear and hierarchical generalized models with slopes as outcomes. *The Journal of Experimental Education*, *73*, 71–95.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York, NY: Springer.

Tsiatis, A. A., & Davidian, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 569–573.

Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, *27*, 4658–4677.

Tu, W., & Zhou, X.-H. (2003). A bootstrap confidence interval procedure for the treatment effect using propensity score subclassification. *Health Services & Outcomes Research Methodology*, *3*, 135–147.

VanderWeele, T. J. (2008). Ignorability and stability assumptions in neighborhood effects research. *Statistics in Medicine*, *27*, 1934–1943.

Verbeek, M. (2004). *A guide to modern econometrics*. West Sussex: John Wiley & Sons.

Waernbaum, I. (2010). Propensity score model specification for estimation of average treatment effects. *Journal of Statistical Planning and Inference*, doi: 10.1016/j.jspi.2010.01.033.

Wainer, H. (2000). The centercept: An estimable and meaningful regression parameter. *Psychological Science*, *11*, 434–436.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

Wang, J., & Donnan, P. (2001). The multiple propensity score for analysis of dose-response relationships in drug safety studies. *Pharmacoepidemiology and Drug Safety*, *10*, 105–111.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. New York, NY: Springer.

Weisberg, H. I. (1979). Statistical adjustments and uncontrolled studies. *Psychological Bulletin*, *86*, 1149–1164.

Weisberg, S. (2005). *Applied linear regression*. Hoboken, NJ: John Wiley & Sons.

West, S. G., & Aiken, L. S. (2005). Multiple linear regression. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1333–1338). Chichester: John Wiley & Sons.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–84). Cambridge, UK: Cambridge University Press.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, *21*, 149–170.

Wilcox, R. R., & Keselman, H. J. (2004). Robust regression methods: Achieving small standard errors when there is heteroscedasticity. *Understanding Statistics in Medicine, 3*, 349–364.

Wold, H. (1956). Causal inference from observational data: A review of end and means. *Journal of the Royal Statistical Society. Series A, 119*, 28–61.

Wolf, A. (2006). *Shorter tests through the adaptive use of planned missing data in sampling designs*. Unpublished doctoral dissertation, Institut of Psychology, Friedrich-Schiller-University of Jena.

Woo, M.-J., Reiter, J. P., & Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in Medicine, 27*, 3805–3816.

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wothke, W. (2000). Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Hillsdale, NJ: Erlbaum.

Yang, L., & Tsiatis, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest-posttest trial. *The American Statistician, 55*, 314–321.

Yanovitzkya, I., Zanutto, E. L., & Hornik, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Evaluation and Program Planning, 28*, 209–220.

Young, S. G., & Bowman, A. W. (1995). Non-parametric analysis of covariance. *Biometrics, 51*, 920–931.

Yuan, K.-H., Marshall, L. L., & Bentler, P. M. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology, 33*, 241–265.

Zanutto, E. L. (2006). A comparison of propensity score and linear regression analysis of complex survey data. *Journal of Data Science, 4*, 67–91.

Zanutto, E. L., Lu, B., & Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics, 30*, 59–73.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software, 11*, 1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software, 16*, 1–16.

Zelen, M. (1990). Randomized conset designs for clinical trials: An update. *Statistics in Medicine, 9*, 645–656.

Appendix
Contents of the Accompanying DVD

The accompanying DVD contains results of the simulation studies presented as additional figures and tables in the digital appendix (`DigitalApendix.pdf`). Moreover, the following documents with supplementary level plots for all conditions of simulation study I and II are provided:

	simulation I	simulation II
Type-I-error rates	`SUPP_I_1.pdf`	`SUPP_II_1.pdf`
Absolute biases of the *ATE*–estimators	`SUPP_I_2.pdf`	`SUPP_II_2.pdf`
Mean squared errors of the *ATE*–estimators	`SUPP_I_3.pdf`	`SUPP_II_3.pdf`
Relative biases of the standard error of the *ATE*–estimators	`SUPP_I_4.pdf`	`SUPP_II_4.pdf`
Mean Squared Errors of the standard error of the *ATE*–estimators	`SUPP_I_5.pdf`	`SUPP_II_5.pdf`
Convergence rates	`SUPP_I_6.pdf`	`SUPP_II_6.pdf`

Erklärung

Die vorliegende Arbeit basiert auf Forschungsarbeiten, die am Lehrstuhl für Methodenlehre und Evaluationsforschung der Friedrich-Schiller-Universität Jena durchgeführt wurden.

In Kenntnis der Promotionsordung der Fakultät für Sozial- und Verhaltenswissenschaften versichere ich ehrenwörtlich, dass ich die von mir eingereichte Dissertation selbst angefertigt und alle Quellen und Hilfsmittel im Text als solche deklariert habe. Ich habe keine Unterstützung eines Promotionsberaters in Anspruch genommen. Über den wissenschaftlichen Austausch mit Kolleginnen und Kollegen hinaus wurde in keiner Phase der Herstellung des Manuskripts entgeltliche oder unentgeltliche Hilfe durch Dritte geleistet. Kein Teil dieser Dissertation wurde an einem anderen Ort für eine wissenschaftliche Qualifikationsleistung eingereicht.

Ich versichere, nach bestem Wissen die Wahrheit gesagt und nichts verschwiegen zu haben.

Jena, den _____ _____

Joachim Ulf Kröhne