

**Average and Conditional Treatment Effects on Count Outcomes:
A Moment-Based Approach**

Von der Philosophischen Fakultät der Rheinisch-Westfälischen Technischen Hochschule
Aachen zur Erlangung des akademischen Grades eines Doktors der Philosophie
genehmigte Dissertation

vorgelegt von

Christoph Kiefer

Berichter: Juniorprofessor Dr. Axel Mayer
Universitätsprofessor Dr. Rolf Steyer

Tag der mündlichen Prüfung: 28. September 2020

“Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online
verfügbar.”

*Wer ihn besitzt, sagt beispielsweise nicht: Hier ist dies oder das
geschehen, wird geschehen, muß geschehen; sondern er erfindet:
Hier könnte, sollte oder müßte geschehn; und wenn man ihm von
irgend etwas erklärt, daß es so sei, wie es sei, dann denkt er:
Nun, es könnte wahrscheinlich auch anders sein.*

Robert Musil, *Der Mann ohne Eigenschaften*

Danksagung

Nacht.

So beginnt Richard Strauß' "Eine Alpensinfonie", eine in Ton gesetzte Gipfelerklimmung. In den vergangenen Monaten habe ich sie oft gehört. Der Anstieg, die blumigen Wiesen, durch Dickicht und Gestrüpp auf Irrwegen – beinahe wie eine Musik gewordene Promotionszeit.

Mein herzlichster Dank geht an Prof. Dr. Axel Mayer, dafür dass er mich auf diese abenteuerreiche Wanderung mitgenommen hat. Ich bin dankbar, dass er mir – auch auf Irrwegen – stets eine selbstständige Erkundung des statistischen Terrains erlaubte und mir Hilfestellung gab, wo immer ich sie benötigte. Ich habe ungeheuer viel gelernt. Danke!

Bemerkenswerterweise ist das Gipfelmotiv der Alpensinfonie bereits früh im Sonnenaufgang zu hören. In diesem Sinne möchte ich auch meinem Zweitgutachter, Prof. Dr. Rolf Steyer sehr herzlich danken, der vor nunmehr fünf Jahren die Anregung zu dieser Bergbesteigung gab und mir schon während der Masterarbeit einen tiefen Einblick in die Irrungen und Wirrungen der psychologischen Methodenlehre gab. Danke!

Wandern – das macht erst gemeinsam recht Spaß. Für stets unterhaltsame Gespräche, entspannende Pausen und gegenseitiges Mutmachen möchte ich mich bei meinen Kolleginnen und Kollegen Aida Alisic, Andre Lerche, Dr. Anja Stertz, Dr. Anna Claus, Benedikt Langenberg, Prof. Dr. Bettina Wiese, Dr. Christian Burk, Julia Norget, Nele Stadtbäumer, Rico Fischer und Ruth Noppeney bedanken. Ich hoffe, sie empfanden diese Wanderung ebenso als bereicherndes Erlebnis. Danke!

Nicht zuletzt gilt mein besonderer Dank auch meiner Frau Mareike, die während meiner Wanderungen oft zu Hause auf meine Wiederkehr wartete. Sie hielt mir stets – und insbesondere in der Schlussphase – den Rücken (und die Hände) frei und lauschte geduldig meinen Berggeschichten. Danke!

Zusammenfassung

In dieser Arbeit wird ein statistisches Modell zur Analyse durchschnittlicher und bedingter Treatment- oder Interventionseffekte auf eine abhängige Zählvariable entwickelt. Die Anwendungsmöglichkeiten eines solchen Modells lassen sich in unterschiedlichen Forschungsbereichen finden: In der Prävention von Drogen- und Substanzmissbrauch lässt sich die Wirkung einer Intervention auf die tägliche oder wöchentliche Anzahl konsumierter alkoholischer Getränke oder Zigaretten untersuchen. In der klinischen Psychologie können Behandlungseffekte auf Zählvariablen wie die Häufigkeit von Heißhungerattacken im Rahmen einer Binge-Eating-Störung oder die Anzahl von Unterrichtsstörungen bei Kindern mit ADHS von Interesse sein. Weitere Beispiele sind die Anzahl richtiger Antworten in einem kognitiven Test, Abwesenheitstage oder auch Häufigkeit der Verletzung von Verkehrsregeln. Das Kernstück unseres *moment-basierten Ansatzes* stellen dabei Regressionen mit logarithmischer Linkfunktion da. Frühere Ansätze zur Schätzung von Treatmenteffekten auf Zählvariablen weisen drei Schwächen auf, die der moment-basierte Ansatz wie folgt angeht:

Erstens, die Definition eines Treatmenteffekts basiert auf der stochastischen Theorie kausaler Effekte (Steyer, Mayer, & Fiege, 2014). Diese Theorie bietet eine eindeutige Definition atomarer bzw. individueller kausaler Effekte als Differenz zwischen bedingten Erwartungswerten der Zählvariablen in einer Behandlungsgruppe und in einer Kontrollgruppe. Zudem werden Kausalitätsbedingungen formuliert, unter welchen durchschnittliche und bedingte Behandlungseffekte als Aggregation atomarer bzw. individueller kausaler Effekte interpretiert werden können. In Regressionsmodellen mit logarithmischer Linkfunktion ist es hingegen üblich die Behandlungseffekte als Verhältnis der bedingten Erwartungen in der Behandlungsgruppe und der Kontrollgruppe zu betrachten. Wir weisen darauf hin, dass diese *Verhältniseffekte* eine andere kausale Bedeutung haben als die *Differenzeffekte*. Beispielsweise ist im randomisierten Experiment das Verhältnis der Gruppenmittelwerte kein Durchschnitt über individuelle Verhältniseffekte. Die Differenz der Gruppenmittelwerte hingegen ist auch der Mittelwert der individuellen Differenzeffekte.

Zweitens, traditionelle Ansätze zur Bestimmung von Treatmenteffekten auf Zählvariablen behandeln die Größe der untersuchten Behandlungsgruppen sowie die Werte der beobachteten Kovariaten als fixiert, d.h. es wird angenommen, dass diese Größen vom Versuchsleiter festgelegt wurden. Falls Gruppengrößen und Kovariaten aber zufällig erhoben werden – was in psychologischen und sozialwissenschaftlichen Studien zumeist der Fall ist –, kann diese Annahme zur Unterschätzung der Standardfehler führen. Eine erhöhte α -Fehlerrate und niedrigere Teststärke für durchschnittliche und bedingte Behandlungseffekte sind die Folge. Der moment-basierte Ansatz bietet hier eine Alternative, indem er sowohl Gruppengrößen als auch die Kovariaten als Zufallsvariablen bzw. stochastisch betrachtet und dadurch die statistische Inferenz verbessert.

Drittens, der moment-basierte Ansatz berücksichtigt Messfehler in den Kovariaten. Während bisherige Ansätze auf einem Generalisierten Linearen Modell mit logarithmis-

cher Linkfunktion (z.B. Poisson-Regressionsmodell) basierten, nutzen wir Strukturgleichungsmodelle. In einem *Negativ-Binomial-Mehrgruppen-Strukturgleichungsmodell* ist es möglich gleichzeitig ein Messmodell für latente Variablen (d.h., Zerlegung messfehlerbehafteter Indikatoren in wahre Werte und Messfehler) und ein Negativ-Binomial-Regressionsmodell mit diesen latenten Variablen als Regressoren zu schätzen. Während die Nichtbeachtung von Messfehlern im Generalisierten Linearen Modell zur verzerrten Schätzung der Regressionskoeffizienten und der darauf aufbauenden Effekte führen kann, werden diese Probleme im moment-basierten Ansatz vermieden.

In dieser Arbeit entwickeln und erweitern wir den moment-basierten Ansatz Schritt für Schritt. Zunächst motivieren wir die Notwendigkeit eines neuen Ansatzes, indem wir die drei zuvor genannten Schwächen der bisherigen Ansätze (d.h. Interpretation exponenzierter Regressionskoeffizienten, marginale Effekte) genauer vorstellen und diskutieren. Danach werden wir die Kernidee des moment-basierten Ansatzes in einem einfachen Fall mit nur einer, aber stochastischen Kovariaten im Generalisierten Linearen Modell einführen. In einer Monte-Carlo-Simulationsstudie vergleichen wir die Güte eines traditionellen Ansatzes (d.h., marginale Effekte) mit dem moment-basierten Ansatz. Drittens übertragen wir den moment-basierten Ansatz auf ein Negativ-Binomial-Multigruppen-Strukturgleichungsmodell und betrachten mehrere, potentiell latente Kovariaten sowie stochastische Gruppengrößen für die Effektschätzung. Viertens lockern wir die Annahme multivariat normalverteilter Kovariaten aus den vorherigen Schritten und entwickeln einen generalisierten moment-basierten Ansatz, der für eine Vielzahl normal- und nicht-normalverteilter Kovariaten eingesetzt werden kann. Damit erreichen wir ein flexibles Modell, das für einen breiten Bereich angewandter Datenanalysen verwendbar ist. Abschließend diskutieren wir Handlungsempfehlungen, wie und wann der moment-basierte Ansatz angewendet werden kann und bieten einen Ausblick auf mögliche Weiterentwicklungen.

Abstract

In this thesis, we develop a statistical framework to investigate average and conditional effects of a treatment or intervention on a count outcome. For example, in the field of drug and substance use prevention, the effects of interventions on the outcome variable daily/weekly count of alcoholic drinks or cigarettes consumed can be examined. In clinical psychology, treatment effects on count outcomes such as instances of binge eating among patients with eating disorders or counts of classroom violations in children with ADHD can be examined. Further examples are treatment effects on counts of correctly answered items in a cognitive test, days of absenteeism, or counts of traffic violations. The framework is termed *moment-based approach* and is based on regression models with a logarithmic link function. The moment-based approach overcomes three major shortcomings of earlier approaches estimating treatment effects on count outcomes:

First, effect definitions are based on the stochastic theory of causal effects (Steyer et al., 2014). This theory provides unambiguous definitions of atomic or individual causal effects as differences between conditional expectations under treatment and under control (i.e., a reference group). Further, causality conditions under which estimated average and conditional treatment effects may be interpreted as aggregates of atomic or individual causal effects are explicated. In contrast, in regression models with a logarithmic link function it is common to inspect treatment effects defined as ratios of conditional expectations under treatment and under control. We caution that these *ratio effects* have a different causal interpretation than *difference effects*. For example, the simple ratio of group averages in a randomized experiment does not represent an average of individual ratio effects – unlike the simple difference of group averages, which is also an average of individual difference effects.

Second, traditional approaches examining treatment effects on count outcomes, treat the size of treatment groups and values of observed covariates as fixed-by-design, that is, predetermined by the experimenter. If in fact group sizes and covariates are not predetermined – as is the case in most psychological and social science studies –, this assumption can lead to underestimated standard errors and, thus, inflated Type 1 error rates and decreased power for average and conditional treatment effect estimates. The moment-based approach offers an alternative by treating both group sizes and covariates as random or stochastic variables, which can substantially improve statistical inferences.

Third, the moment-based approach allows to account for measurement error in covariates. While earlier approaches were based on generalized linear models with a logarithmic link function (e.g., Poisson regression models), we use a structural equation modeling framework. In a *negative binomial multigroup structural equation model*, it is possible to simultaneously estimate measurement models for latent variables (i.e., decomposing true scores and measurement error in fallible indicators) and a negative binomial regression model with these latent variables as regressors. While ignoring measurement error in the generalized linear models might lead to biased estimation of both the regression coefficients

and the corresponding effect estimates, these issues are circumvented in the moment-based approach.

In this thesis, the moment-based approach is developed and extended in a step-by-step manner. First, we motivate the need for a new approach by presenting and discussing the three aforementioned shortcomings of earlier approaches (i.e., interpretation of exponentiated regression coefficients and marginal effects) in detail. Second, we introduce the key idea of the moment-based approach in a simple case considering only a single, but stochastic covariate using a generalized linear model. In a Monte Carlo simulation study, we compare the performance of a traditional (marginal effects) approach with the performance of the moment-based approach. Third, we extend the statistical framework to the negative binomial multigroup structural equation model and include multiple, possibly latent covariates as well as stochastic group sizes in our effect computations. Fourth, we relax the assumption of multivariate normally distributed covariates from previous steps and derive a generalized moment-based approach accounting for various kinds of normally and non-normally distributed covariates. As a result, we achieve a suitable approach for a broader range of applied data analyses. Finally, we discuss guidelines how and when the moment-based approach should be applied and have an outlook on possible further developments.

Contents

List of Figures	i
------------------------	----------

List of Tables	iii
-----------------------	------------

1 Introduction	1
1.1 Effects and Causality	2
1.1.1 A Primer on Terminology and Notation	3
1.1.2 Single-Unit Trial	4
1.1.3 Average and Conditional Causal Effects	4
1.1.4 Causality Conditions	6
1.1.5 Excursus: Effects Defined as Ratios	8
1.2 Causal Inference for Count Outcomes	10
1.2.1 Regression-Based Causal Inference	11
1.2.2 Propensity Score-Based Causal Inference	14
1.2.3 Double Robust Methods	16
1.3 Statistical Frameworks for Count Outcomes	16
1.3.1 Family of Poisson Regression Models	17
1.3.2 Semiparametric Count Regression and M-Estimation	18
1.3.3 Structural Equation Modeling	19
1.3.4 Bayesian Count Regression Models	19
1.4 Goals and Structure of this Thesis	20
2 A New Approach with Stochastic Covariates	23
2.1 Introduction	24
2.1.1 Definition and Terminology of Effects	26
2.2 Regression with Log Link	27
2.2.1 Treatment Effects in Regressions with Log Link	27
2.3 Estimation of Regression Coefficients	28
2.4 Estimates and Standard Errors of Treatment Effects	29
2.4.1 Empirical-Distribution-Based Approach	29
2.4.2 Moment-Based Approach with Stochastic Covariates	30
2.5 Simulation Study	33
2.5.1 Data Analysis	34
2.5.2 Results	35
2.6 Empirical Illustration	37
2.7 Discussion	41
Appendix	44
2.A Non-Collapsibility of Ratio Effects	44
2.B Standard Error of Moment-Based Estimator	45
2.C Supplementary Material: Simulation Details	50
2.D Supplementary Material: CountEffects Package	57

3	Accounting for Latent Covariates	59
3.1	Introduction	60
3.2	Definition and Terminology of Treatment Effects	62
3.2.1	Conditional Effects and the CE Function	62
3.2.2	Average Effect and Effects at the Mean	64
3.2.3	Excursus: Effects Defined as Ratio	65
3.3	Negative Binomial Multigroup SEM (NB-MG-SEM)	66
3.4	Extended Moment-Based Approach	68
3.4.1	Conditional Effects and CE Function	68
3.4.2	Average Effect and Effect at the Mean	68
3.5	Illustrative Example	69
3.5.1	Measures	70
3.5.2	Model	71
3.5.3	Average and Conditional Treatment Effects	72
3.5.4	Comparative Analyses	73
3.5.5	Results	74
3.6	Summary and Conclusions	78
3.6.1	Limitations and Further Research	79
	Appendix	81
3.A	Estimation	81
3.B	Regression Coefficients Transformation	83
3.C	Supplementary Material: Shiny App on Attenuation Bias	84
3.D	Supplementary Material: Mplus Syntax for Illustrative Example	87
4	Accounting for Non-Normal Covariates	91
4.1	Introduction	92
4.2	Definition and Terminology of Treatment Effects	93
4.2.1	Conditional Effects Function	94
4.2.2	Average Treatment Effect	95
4.2.3	Conditional Effects given X and K	95
4.2.4	Regression Parameterization and Covariates Factorization	96
4.3	Moment-Based Approach Assuming Multivariate Normality	97
4.4	Accounting for Non-Normal Continuous Covariates	98
4.4.1	Case 1: Known Non-Normal Joint Distribution	98
4.4.2	Case 2: Factorization of Joint Distribution	99
4.4.3	Case 3: Approximation with Finite Gaussian Mixtures	100
4.5	Negative Binomial Multigroup Structural Equation Model	100
4.6	Illustrative Example	101
4.6.1	Sample	102
4.6.2	Measures	102
4.6.3	Model	103
4.6.4	Average and Conditional Treatment Effects	103
4.6.5	Results	105
4.7	Summary and Conclusions	109
4.7.1	Limitations and Further Research	110
	Appendix	113
4.A	Factorization of the Joint Distribution	113
4.B	Model Estimation	114
4.C	Effect Estimation	116

4.D	Supplementary Material: R Code for Model Estimation	119
4.E	Supplementary Material: R Code for Effect Estimation	121
5	General Discussion	125
5.1	General Overview and Summary	125
5.1.1	Treatment Effect Definition	125
5.1.2	Stochastic Covariates and Group Sizes	126
5.1.3	Latent Covariates	127
5.1.4	Non-Normal Covariates	128
5.1.5	Guidelines for Use of Moment-Based Approach	129
5.2	Limitations and Future Research	130
5.2.1	In-Depth Comparison of Effect Estimation Approaches	130
5.2.2	Non-Canonical Use of Link Functions	131
5.2.3	Zero-Inflated Count Data and Hurdle Models	132
5.2.4	Marginal Effects for Latent Covariates Revisited	133
5.2.5	Treatment Effects Defined as Ratios	134
5.3	Concluding Remarks	134
	References	135

List of Figures

2.1	Plots for our illustrative example showing the group-specific regressions of Y on Z	31
2.2	Bias $B(\widehat{ATE})$ for the estimators $\widehat{ATE}_{EMP}$ (ATE EMP), $\widehat{ATE}_{MOM}$ (ATE MOM), and $\widehat{TEMZ}$ (TEMZ).	36
2.3	Coverage of 95% confidence intervals based on the respective standard error estimates SE_{EMP} (SE EMP) and SE_{MOM} (SE MOM).	38
2.4	Plots of Type 1 error rates and power for standard error $\widehat{ATE}_{EMP}$ (SE EMP) and standard error for $\widehat{ATE}_{MOM}$ (SE MOM).	39
3.1	Illustration of the incident rate ratio with and without interaction effect	66
3.2	Path diagram for the illustrative example	73
4.1	Path diagram for the illustrative example	104

List of Tables

1.1	Joe and Ann With Randomized Assignment: Conditional Expectation and Their Corresponding Ratio	10
2.1	Illustrative example of how the ATE is computed using the empirical approach.	30
2.2	Moment-generating functions for several unconditional distributions of Z	32
2.3	Design of the simulation study.	34
3.1	Assumptions within the extended moment-based approach	63
3.2	Maximum likelihood estimates for the measurement model for the three latent covariates	75
3.3	Maximum likelihood estimates for the two-group structural equation model with stochastic group sizes and logarithmic link function	76
3.4	Estimated average and conditional effects of therapy ($X = 1$ vs. $X = 0$) compared to the control group in the simulated data example.	77
3.5	Comparative Analyses	78
4.1	Maximum likelihood results for group-specific model parameters	106
4.2	Estimated average and conditional effects of memory training ($X = 1$ vs. $X = 0$), reasoning training ($X = 2$ vs. $X = 0$), and speed of processing training ($X = 3$ vs. $X = 0$) compared to the control group	108

1 Introduction

This thesis is concerned with causal inference on count outcome variables – more precisely, with statistical inference about average and conditional total effects estimated using regression models with a logarithmic link function. In social and health sciences, the effectiveness of a treatment or an intervention is often evaluated on a count outcome variable. For example, in the field of drug and substance use prevention, the effects of interventions on the outcome variable daily/weekly count of alcoholic drinks or cigarettes consumed have been examined (Garrett et al., 2018; Layland, Calhoun, Russell, & Maggs, 2018; Muench et al., 2017; Neighbors et al., 2015). In clinical psychology, treatment effects on count outcomes such as instances of binge eating among patients with eating disorders (Kenny, Van Wijk, Singleton, & Carter, 2018) or count of classroom violations in children with ADHD (Helseth et al., 2015; Pelham, Jr. et al., 2016) have been examined. In medicine, the effectiveness with which the count of MRI lesions and relapses in multiple sclerosis patients can be reduced was examined (Cascione et al., 2018; Cree et al., 2018). Further examples are treatment effects on counts of correctly-answered items in a cognitive test (Macdonald, Germine, Anderson, Christodoulou, & McGrath, 2017; Nusser & Weinert, 2017), days of absenteeism (Schaumberg & Flynn, 2017), or counts of traffic violations (Precht, Keinath, & Krems, 2017).

Commonly, a generalized linear model (GLM) with a logarithmic link function is applied to regress the count outcome variable on the treatment variable and several covariates. It consists of a linear function of regressors (systematic component) linked to the logarithm of the conditional expectation of the count outcome variable. Hereby, the natural lower bound of a count at zero is incorporated. Depending on the conditional distribution of the count outcome (random component), the statistical model is referred to as *Poisson regression*, *quasi-Poisson regression*, or *negative binomial regression* (Hilbe, 2007; McCullagh & Nelder, 1996). All three count regression models come with a certain amount of heteroscedasticity. While the conditional Poisson distribution yields *equidispersion* (i.e., conditional expectation and conditional variance of the count outcome are equal), the quasi-Poisson and negative binomial regression allow for *overdispersion* (i.e., conditional variance of the count outcome can be greater than its conditional expectation). In some versions of negative binomial regression, *underdispersion* (i.e., conditional variance of the count outcome is smaller than its conditional expectation) is possible. For an overview of variance parameterizations in count regression models, see Hilbe (2011, ch. 10).

It is important to note that the coefficients of regressions with a logarithmic link function, like those of any nonlinear regression model, do not correspond directly to (average)

treatment effects on the original count scale of the outcome (Greene, 2007, p. 729). Therefore, it is often necessary to transform the regression coefficients to the original count scale, for example, by exponentiating the coefficients (Coxe, West, & Aiken, 2009), which yields a ratio of conditional expectations of the count outcome under treatment and under control. An alternative are marginal or partial effects (Hilbe, 2011), which enable computing a discrete change in the count outcome due to the treatment compared with a reference group. We provide an overview of methods for causal inference on count outcomes in Section 1.2.

However, these traditional approaches have three major shortcomings, which we will describe in more detail in the following sections: First, effects defined as ratios suffer from some undesirable properties, such as non-collapsibility (Greenland, 1996). That is, even in absence of confounding causes (e.g., in a randomized controlled trial), a ratio effect is prone to change due to controlling for additional covariates, making it difficult to identify a *true* ratio effect. Second, covariates' values are considered to be fixed-by-design, for example, in the marginal effects approach. It has been argued before, that this assumption can lead to underestimated standard errors and therefore inflated Type 1 error rates and decreased power (Chen, 2006; Kröhne, 2009; Y. Liu, West, Levy, & Aiken, 2017). Related approaches such as g-computation (J. Robins, 1986) do not offer analytical standard errors, but rely on bootstrap methods (Daniel, Stavola, & Cousens, 2011; Wang, Nianogo, & Arah, 2017). Third, in traditional approaches the covariates are often assumed to be measured without error. Nevertheless, in psychological studies researchers often face so-called *latent* constructs like empathy or intelligence, which cannot directly be observed, but only with fallible indicators (e.g., items in a questionnaire). Neglecting the measurement error in covariates can lead to biased effect and standard error estimates (Mayer, 2019; Sengewald, Steiner, & Pohl, 2018). The goal of the present thesis is to address these issues for treatment effect estimation in count regression models with a logarithmic link function.

1.1 Effects and Causality

The concept of causal inference, definitions of causal effects, and techniques of deriving estimable quantities have a long tradition. There is the *Neyman-Rubin tradition* (Splawa-Neyman, 1923/1990) with the definition of individual, conditional, and average treatment effects leading to Rubin's causal model (Rubin, 2005) based on potential or counterfactual outcomes. Counterfactual outcomes are unobserved outcomes, that would have been observed under different conditions. For instance, if we observe a participant's count of correctly answered items in a cognitive test after a certain treatment, a counterfactual outcome would be the count of correctly answered items if the particular participant had not received the treatment. A counterfactual outcome cannot be observed, yet can be estimated under certain assumptions. The *Campbellian tradition* (Shadish, Cook, & Campbell, 2002) has a stronger focus on questions and problems in applied scenarios, as

in improperly carried out randomization or unobserved values of the outcome variable for some participants (see also West, Cham, & Liu, 2014). Another tradition was initiated by Pearl (2009) using *graphical modeling*.

In this thesis, the stochastic theory of causal effects (Steyer, Gabler, von Davier, Nachtingall, & Buhl, 2000; Steyer, Mayer, & Fiege, 2014) serves as a conceptual framework for defining individual, conditional and average causal effects. In recent presentations (Mayer, 2019; Steyer et al., 2014), the stochastic theory of causal effects is described as a stochastic process. On the one hand, it considers the point in time at which a random variable occurs for the first time. Thereby, it distinguishes between *covariates* (prior or simultaneous to treatment) and *mediators* (after treatment) and derives causality conditions for total, direct, and indirect effects of the treatment. On the other hand, its concepts and definitions are grounded in probability theory and therefore allow for an integration of stochasticity within the effect computation. For example, effect definitions are compatible with a theory of latent variables (Mayer, 2019; Sengewald et al., 2018) and hereby allow to account for measurement error in fallible covariates. The definitions of causal effects and causality conditions in this chapter follow the presentation of Steyer et al. (2014).

1.1.1 A Primer on Terminology and Notation

The notation and terminology in the following sections and chapters follows Steyer et al. (2014) and Mayer (2019). Throughout this thesis, however, the probabilistic foundation of the stochastic theory of causal effects is presented in a less formal notation. For the formal notation and the corresponding definitions and theorems, we refer to Steyer et al. (2014) and Steyer and Nagel (2017). To mention one, we often use the partial conditional expectation $E(Y|X=x, Z)$ which corresponds to

$$E(Y|X=x, Z) \underset{p_{X=x}}{=} E^{X=x}(Y|Z).$$

For assumptions and definitions of $E(Y|X=x, Z)$ and $E^{X=x}(Y|Z)$, see Definitions 14.29 and 14.7 (or Remark 14.18 respectively), for a theorem on their relation, see Theorem 14.33 in Steyer and Nagel (2017). Additionally, whenever an equivalence with respect to a measure p is stated, we abbreviate $\underset{p}{=}$ with $=$, but still mean the terms to be only *almost surely* equal. In case conditional expectations or probabilities involve $(X=x)$, the statements are meant to apply for all x .

In contrast to Steyer et al. (2014), we do not discuss the role of mediating variables and consequently do not distinguish between direct and indirect effects of the treatment, but exclusively consider total effects. Accordingly, when in the following sections and chapters we speak of average effects, causal effects, and suchlike, we always mean average *total* effects, causal *total* effects, et cetera. Further, covariates are defined as variables that occur before or simultaneously to the treatment variable, but *not* after the treatment. Variables that occur after the treatment are considered mediator variables and are not discussed in

this thesis. For more information on causal mediation analysis with count mediators or outcomes, see [Cheng et al. \(2017\)](#), [Guo, Small, Gansky, and Cheng \(2017\)](#), and [O’Rourke and Vazquez \(2019\)](#). For a more general overview of causal inference considering mediating variables and corresponding direct and indirect effects, see [Mayer, Thoemmes, Rose, Steyer, and West \(2014\)](#) and [J. M. Albert and Nelson \(2011\)](#).

1.1.2 Single-Unit Trial

In the stochastic theory of causal effects, the definition of causal effects as well as the derivation of causality conditions are based on a random experiment, called *single-unit trial*. In this thesis, we confine ourselves to the case in which we consider a count variable Y with values $y \in \mathbb{N}_0$ and a discrete treatment variable X with $p+1$ levels and $P(X=x) > 0, \forall x=0, 1, \dots, p$. For simplicity, we always treat $X=0$ as reference or control group. Analogous to [Steyer et al. \(2014\)](#), the following single-unit trial is considered:

- (a) sample a person u out of a set of persons Ω_U
- (b) observe the person’s value z of a fallible (multivariate) covariate Z (i.e., prior or simultaneous to treatment)
- (c) assign the person to one of several treatment conditions (represented by the values x of the treatment variable X) or observe its (possibly self-selected) assignment
- (d) observe the person’s value y of the outcome variable Y

The single-unit trial describes the empirical phenomenon which is typically investigated in the causal inference literature both for experimental and quasi-experimental settings. It does *not* represent a sampling process, but the random experiment of drawing a single person etc. In the same vein, the causal effects in the next section are defined with regard to the single-unit trial.

1.1.3 Average and Conditional Causal Effects

Based on the single-unit trial, we can define a *total-effect true outcome* variable

$$\tau_x := E(Y|X=x, C_X),$$

where C_X is a *comprehensive potential confounder* variable. The comprehensive potential confounder comprises all variables that influence choice or outcome of a treatment variable. The true outcome variable τ_x represents the expected count of Y for a person u , if he or she was assigned to treatment condition x controlling for further variables affecting the outcome or choice of treatment beyond the person variable U . For example, if a person was assigned to psychotherapy treating substance use disorder $X=t$, τ_t reflects the expected weekly count of consumed alcoholic drinks after receiving psychotherapy. More formally, τ_x

represents the true expected outcome for keeping X constant on its value x and controlling for C_X .

The atomic treatment effect variable δ_{t0} of treatment $X = t$ versus control $X = 0$ is then defined as:

$$\delta_{t0} = \tau_t - \tau_0.$$

The treatment effect variable δ_{t0} reflects the difference between the true expected count outcome if receiving treatment (i.e., τ_t) versus the true expected count outcome if not being treated (i.e., τ_0).

If we aggregate the atomic effects to the level of the person variable U , then

$$E(\tau_t|U) - E(\tau_0|U)$$

is called *individual treatment effect* variable. Continuing our example of examining treatment effects of psychotherapy on substance use disorder, $E(\tau_t|U)$ reflects a person's expected count of alcoholic drinks after receiving psychotherapy and $E(\tau_0|U)$ reflects the person's expected count of alcoholic drinks if not treated. Hence, the difference is the treatment effect for that person. In an empirical study, it is usually not possible to estimate atomic or individual treatment effects, because a person can only be observed under one of the treatment conditions. This is known as the *fundamental problem of causal inference* (Holland, 1986). Nevertheless, Mayer (2019) refers to two strong assumptions under which individual treatment effects are estimable. First, *conditional unit-treatment homogeneity* (Steyer, Gabler, von Davier, & Nachtigall, 2000) assuming $U = C_X$ and second, *no pre-post change in the control group* (Steyer, 2005).

It is, however, possible to estimate aggregates of the atomic treatment effects under weaker causality conditions. Given the aforementioned single-unit trial, the starting point for these aggregations is the conditional expectation $E(Y|X = x, Z)$ and the conditional effect function

$$g_t(Z) := E(Y|X = t, Z) - E(Y|X = 0, Z).$$

In contrast to the atomic treatment effect, the conditional effect function does neither contain the comprehensive potential confounder C_X nor the person variable U , but only the covariate Z . Thus, it does not necessarily yield a causal, yet rather a *prima facie* effect. If the conditional effect function is causally unbiased, that is,

$$g_t(Z) = E(\delta_{t0}|Z),$$

then $g_t(Z)$ can be used to identify average and conditional causal effects. That is to say if unbiasedness of the conditional effect function holds, the average treatment effect is given

as a unconditional expectation over atomic or conditional effects:

$$\begin{aligned} ATE_{t0} &:= E(\delta_{t0}) \\ &= E[g_t(Z)]. \end{aligned}$$

Usually, whether the conditional effect function $g_t(Z)$ is causally unbiased is not testable, because δ_{t0} cannot be observed or estimated directly. However, there are some causality conditions which imply unbiasedness of the conditional effect function $g_t(Z)$ or the conditional expectations $E(Y|X = x, Z)$ respectively, some of which are even testable in empirical settings. We will present and discuss them in the next section.

1.1.4 Causality Conditions

In this section, we will present two kinds of causality conditions in detail: *independent cause conditions* and *regressively independent outcome conditions* (Mayer, Thoemmes, et al., 2014). Both conditions are used later in this thesis. For an overview of further causality conditions, we refer to Mayer, Thoemmes, et al. (2014), Steyer, Gabler, von Davier, and Nachtigall (2000), and Steyer et al. (2014).

Unbiasedness of Conditional Expectations

In the previous section, we introduced unbiasedness of the conditional effect function $g_t(Z)$, which is obtained if the conditional expectations $E(Y|X = x, Z)$ are causally unbiased. The following causality conditions either implicate unbiasedness of $E(Y|X = x)$, which is given if

$$E(Y|X = x) = E(\tau_x)$$

holds, or, alternatively, $E(Y|X = x, Z)$, which is given if

$$E(Y|X = x, Z) = E(\tau_x|Z).$$

holds. In summary, unbiasedness ensures that a conditional expectation represents an aggregation of the true outcome variable. For example, in a non-randomized experiment, the simple group mean $E(Y|X = 1)$ is not necessarily an unbiased aggregate of the true outcome variable τ_1 and thus cannot straightforwardly be used for causal inferences. Unbiasedness of $E(Y|X = x)$ or $E(Y|X = x, Z)$ cannot be tested directly, but is implied by several causality conditions which – in turn – are testable. In the following presentation of causality conditions, we will always indicate for which conditional expectation unbiasedness is implied.

Independent Cause Conditions

Randomization is often described as the gold standard to achieve causally unbiased effect estimates. Individuals are randomly assigned to one of the treatment conditions, so independence of the treatment variable X and a comprehensive potential confounder C_X is obtained. Formally, the condition can be denoted as:

$$P(X = x) = P(X = x|C_X) \Rightarrow E(Y|X = x) = E(\tau_x) \quad (1.1)$$

The probability to receive treatment $X = x$ is independent from all potentially confounding variables, which implies unbiasedness of $E(Y|X = x)$. This causality condition is used for effect estimation in the illustrative examples of Chapters 2, 3, and 4.

The unconditional independent cause condition is testable, because Equation (1.1) implies

$$P(X = x|Z) = P(X = x),$$

for any covariate Z . Hence, if a covariate Z can be found predicting the treatment probability, Equation (1.1) does not hold.

Randomized treatment assignment conditional on a covariate also is an independent cause condition. This causality condition, termed Z -conditional independence of X and the comprehensive potential confounder C_X , is given by

$$P(X = x|Z) = P(X = x|C_X) \Rightarrow E(Y|X = x, Z) = E(\tau_x|Z) \quad (1.2)$$

and can be created by covariate selection, that is, by choosing Z for which the condition holds. Analog to the unconditional independence of X and C_X , this causality condition can be falsified in an empirical setting, if a covariate W can be identified for which

$$P(X = x|Z) \neq P(X = x|Z, W)$$

does not hold.

Regressively Independent Outcome Condition

Another condition implying unbiasedness is unconditional regressive independence of Y from the comprehensive potential confounder C_X :

$$E(Y|X = x) = E(Y|X = x, C_X) \Rightarrow E(Y|X = x) = E(\tau_x).$$

However, this causality condition can neither be created by randomization nor by covariate selection. It basically means, the expected outcome of Y depends on the treatment variable

alone. In this sense, it can be empirically falsified by a covariate Z for which

$$E(Y|X=x) = E(Y|X=x, Z)$$

does not hold.

More interesting, from a methodological point of view, is the conditional regressive independence of Y from all covariates of X given X and Z , or formally:

$$E(Y|X=x, Z) = E(Y|X=x, C_X) \Rightarrow E(Y|X=x, Z) = E[\tau_x|Z]. \quad (1.3)$$

Here, a comprehensive potential confounder C_X does not contain more information than Z . This causality condition is used in the simulation study of Chapter 2.

The conditional regressive independence condition is also empirically testable and can be used for covariate selection. If Equation (1.3) holds, for any additionally observed covariate W

$$E(Y|X=x, Z) = E(Y|X=x, Z, W)$$

is implied. Hence, if a covariate W can be found which predicts Y beyond X and Z , Equation (1.3) does not hold.

These causality conditions give us an idea how causal unbiasedness of the conditional expectations $E(Y|X=x)$ and $E(Y|X=x, Z)$ can be achieved. In non-randomized, observational studies, it is common to apply covariate selection strategies aiming at either the conditional independent cause condition given in Equation (1.2) or the conditional regressively independent outcome condition given in Equation (1.3). In Section 1.2, we will discuss different approaches of using covariate selection for causal inference.

1.1.5 Excursus: Effects Defined as Ratios

In the stochastic theory of causal effects, the atomic treatment effect δ_{t0} is defined as *difference* between the true outcome variables τ_t and τ_0 . Nevertheless, treatment effects on count outcomes are often reported as *ratios* of conditional expectations. We will present and discuss the traditional approach of obtaining a ratio effect in Section 1.2.1. In this excursus, we will briefly consider some properties of treatment effects defined as ratios from the perspective of the stochastic theory of causal effects. Further, we will illustrate how different approaches to aggregation of ratios can lead to contradictory results. For an introduction to the difficulties of defining a *true ratio effect*, see [Maldonado and Greenland \(2002\)](#).

Definition and Properties of Ratio Effects

Analog to the definition of the atomic treatment effect δ_{t0} , the atomic ratio treatment effect ρ_{t0} can be defined as:

$$\rho_{t0} := \frac{\tau_t}{\tau_0}.$$

While for treatment effects defined as differences both aggregation over the conditional expectations τ_x and aggregation over the atomic treatment effect d_{t0} yield the same result, this is not the case for ratio effects. For example, the average treatment effect is

$$ATE_{t0} = E(\tau_t - \tau_0) = E(\tau_t) - E(\tau_0),$$

but for an “average” ratio effect these aggregations yield two generally non-equivalent quantities, namely:

$$ARTE_{t0} := E \left[\frac{\tau_t}{\tau_0} \right] \neq \frac{E[\tau_t]}{E[\tau_0]} =: RUC_{t0}.$$

We call $ARTE_{t0}$ the *average over atomic ratio treatment effects* and RUC_{t0} the *ratio of unconditional expectations of the true outcome variables*. It can easily be seen, that the aforementioned causality conditions can be used to identify the components of the RUC_{t0} as causally unbiased. That is, if a corresponding causality condition holds, the ratio of unconditional expectations is causally unbiased. Yet, the RUC_{t0} cannot be interpreted as an average over atomic ratio treatment effects, but as ratio of causally unbiased expectations of the true outcome variables. In magnitude and direction it is indicative for magnitude and direction of the ATE_{t0} , because they have the same constituting parts (i.e., $E[\tau_x]$):

$$RUC_{t0} = 1 \Leftrightarrow ATE_{t0} = 0$$

$$RUC_{t0} > 1 \Leftrightarrow ATE_{t0} > 0$$

$$RUC_{t0} < 1 \Leftrightarrow ATE_{t0} < 0.$$

In contrast, the average ratio treatment effect $ARTE_{t0}$ is non-collapsible, which means, that independent cause conditions (e.g., randomization) do not imply unbiasedness of the $ARTE_{t0}$, for example:

$$P(X = x) = P(X = x, C_X) \not\Rightarrow \frac{E(Y|X=t)}{E(Y|X=0)} = E \left[\frac{E(Y|X=t, C_X)}{E(Y|X=0, C_X)} \right].$$

This is due to a change of the scale before the aggregation, as a ratio does not directly correspond to the level of the outcome variables (i.e., counts). As a consequence, a simple ratio of group averages would not suffice as causal estimate of the $ARTE_{t0}$ in randomized experiments. Hence, some kind of covariate-adjustment would be necessary despite of randomization. This issue has previously been discussed, for example, by [Greenland](#),

Table 1.1: Joe and Ann With Randomized Assignment: Conditional Expectation and Their Corresponding Ratio

u	$E(Y X=0, U=u)$	$E(Y X=1, U=u)$	$\frac{E(Y X=1, U=u)}{E(Y X=0, U=u)}$	$P(U=u)$	$P(X=x U=u)$
Joe	8	5	0.625	.5	.5
Ann	4	6	1.500	.5	.5

Robins, and Pearl (1999).

Example: Joe and Ann with Ratio Effects

The average ratio treatment effect $ARTE_{t0}$ and the ratio of unconditional expectations RUC_{t0} can yield substantially different results. Table 1.1 provides an example of a simple random experiment: Sample a person from the set $\Omega_U := \{Joe, Ann\}$, randomly assign the sampled person to the control group $X=0$ or the treatment group $X=1$ and observe the count outcome Y . The corresponding $ARTE_{t0}$ and RUC_{t0} for this example are:

$$ARTE_{t0} = E \left[\frac{E(Y|X=1, U)}{E(Y|X=0, U)} \right] = 0.625 \cdot 0.5 + 1.500 \cdot 0.5 = 1.0625$$

$$RUC_{t0} = \frac{E[E(Y|X=1, U)]}{E[E(Y|X=0, U)]} = \frac{5 \cdot 0.5 + 6 \cdot 0.5}{8 \cdot 0.5 + 4 \cdot 0.5} = 0.91\bar{6}.$$

While the $ARTE_{t0}$ shows an expected increase of about 6% in the outcome, the RUC_{t0} yields a 8 % decrease in the treatment group compared to the control group.

Again, we caution the use of effects defined as ratios, because aggregates over atomic ratio effects are not collapsible and the simple ratio of group averages in a randomized experiment may be causally unbiased, but not an “average ratio”. Researchers ought to be aware of these limitations and to us (causal) interpretation seems to be more straightforward for effects defined as differences.

1.2 Causal Inference for Count Outcomes

In the previous section, we gave non-parametric effect definitions and the conditional effect function $g_t(Z)$ based on the conditional expectation $E(Y|X=x, Z)$. We also presented several causality conditions which can be achieved using careful covariate selection. However, as Schafer and Kang (2008, p. 280) state: “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.”

In this section, we will present and discuss some of these methods and solutions which can roughly be categorized as (a) regression-based methods aiming at the conditional regressively independent outcome condition given in Equation (1.3), (b) propensity score-based methods aiming at the conditional independent cause condition given in Equation (1.2), and (c) a combination of both regression- and propensity score-based methods (also called double robust methods).

The regression-based methods are of particular interest when examining count outcomes, because they focus on the (potentially non-linear) relationship between the count outcome, the treatment variable, and covariates. In contrast, the propensity score-based methods focus on the relationship between the treatment variable and the covariates and thus are less specific for count data than regression-based methods. The moment-based approach introduced and developed in this thesis can be classified as a regression-based method. The following subsections emphasize regression-based methods, exponentiated regression coefficients and the marginal effects approach in particular, because they are used as reference and for comparison throughout the development of the moment-based approach. Nevertheless, we will briefly describe propensity score- and double robust methods and (if existent) their specific solutions for count outcomes.

1.2.1 Regression-Based Causal Inference

Regression-based methods usually aim to provide causal inferences by selecting covariates that satisfy the causality condition of conditional regressively independence, that is,

$$E(Y|X = x, Z) = E(Y|X = x, C_X),$$

but they may also be applied in randomized experiments for examining differential or conditional effects respectively.

For a count outcome variable Y , the conditional expectation of Y given X and Z can always be expressed as:

$$\begin{aligned} E(Y|X, Z) &= \exp[f(X, Z)] \\ \log[E(Y|X, Z)] &= f(X, Z), \end{aligned}$$

where the logarithm of the conditional expectation $E(Y|X = x, Z)$ is given by the predictor function $f(X, Z)$. Hence, such regressions are called *regressions with a logarithmic link function*. In this thesis, we assume $f(X, Z)$ to be a linear function of X , Z , and their interaction. For a treatment variable with $p + 1$ levels and a M -dimensional covariate $\mathbf{z} = (Z_1, Z_2, \dots, Z_M)$, the function $f(X, Z)$ is given by:

$$f(X, Z) = \gamma_{00} + \sum_{x=1}^p \gamma_{x0} \cdot I_{X=x} + \sum_{m=1}^M \gamma_{0m} \cdot Z_m + \sum_{x=1}^p \sum_{m=1}^M \gamma_{xm} \cdot I_{X=x} \cdot Z_m.$$

For example, let us assume a binary treatment variable ($X=0$ vs. $X=1$) and a single covariate Z_1 , then the regression with a logarithmic link function is:

$$\begin{aligned} E(Y|X, Z_1) &= \exp(\gamma_{00} + \gamma_{10}X + \gamma_{01}Z_1 + \gamma_{11}XZ_1) \\ \log[E(Y|X, Z_1)] &= \gamma_{00} + \gamma_{10}X + \gamma_{01}Z_1 + \gamma_{11}XZ_1. \end{aligned} \quad (1.4)$$

This parameterization is – apart from the logarithmic link function – similar to linear regression parameterization for effect computation, for example, in the EffectLiteR approach (Mayer, Dietzfelbinger, Rosseel, & Steyer, 2016). It is also analogous to the parameterization of the well-known Poisson distribution, which provides one possible statistical framework for estimating the regression coefficients γ_{xm} .

Exponentiated Regression Coefficient

In absence of an interaction effect (i.e., $\gamma_{11} = 0$), the most common approach to evaluate the effectiveness of treatment X is exponentiation of the regression coefficient γ_{10} :

$$\exp(\gamma_{10}) = \frac{E(Y|X=1, Z_1 = z_1)}{E(Y|X=0, Z_1 = z_1)},$$

which is constant for all values z_1 . It reflects the ratio of expected counts under treatment versus under control at any given value of the covariate. If we additionally assume a randomized experiment, the exponentiated coefficient $\exp(\gamma_{10})$ yields the ratio of unconditional expectation of the true outcome variables RUC_{10} . In Section 1.1.5, we discussed that the RUC_{10} is not an aggregation of atomic ratio treatment effects.

Further, if we introduce an interaction effect into the model (i.e., $\gamma_{11} \neq 0$), the interpretation of $\exp(\gamma_{10})$ gets limited, because it only reflects the ratio for a zero value of the covariate:

$$\exp(\gamma_{10}) = \frac{E(Y|X=1, Z_1=0)}{E(Y|X=0, Z_1=0)}.$$

In Chapter 3, we show that this ratio varies for different values z_1 and that mean-centering does not necessarily yield a meaningful quantity, because the ratio at the mean of the covariate is neither equivalent to the RUC_{10} nor to the $ARTE_{10}$.

Marginal Effects

Marginal effects offer a way to estimate the conditional effect function $g_t(Z)$ based on a Poisson regression model. In the aforementioned example, the conditional effect function $g_1(Z_1)$ is given by:

$$\begin{aligned} g_1(Z_1) &= E(Y|X=1, Z_1) - E(Y|X=0, Z_1) \\ &= \exp(\gamma_{00} + \gamma_{10} + \gamma_{01}Z_1 + \gamma_{11}Z_1) - \exp(\gamma_{00} + \gamma_{01}Z_1). \end{aligned}$$

For each observation i , the estimated effect function $\hat{g}_1(Z_1)$ is evaluated at the covariate value z_{1i} resulting in an estimated treatment effect for each observation. These are called marginal effects, because they are computed over the marginal empirical distribution of the covariate. Marginal effect estimation is usually carried out as a two step procedure. First, a statistical model from the family of Poisson regression models is estimated. For our simple example with count outcome Y , binary treatment variable X , and one-dimensional covariate Z_1 , the corresponding Poisson model equation is given in Equation (1.5). The first step yields estimates $\hat{\gamma}_{00}$, $\hat{\gamma}_{10}$, $\hat{\gamma}_{01}$, and $\hat{\gamma}_{11}$ for the regression coefficients. Second, the regression coefficient estimates are used to generate an estimated conditional effect function $\hat{g}_1(z_{1i})$:

$$\hat{g}_1(z_{1i}) = \exp(\hat{\gamma}_{00} + \hat{\gamma}_{10} + \hat{\gamma}_{01}z_{1i} + \hat{\gamma}_{11}z_{1i}) - \exp(\hat{\gamma}_{00} + \hat{\gamma}_{01}z_{1i}).$$

Aggregates of these conditional effects estimates can be obtained. For example, the estimated average treatment effect $\widehat{ATE}_{10}$ is given by:

$$\widehat{ATE}_{10} = \frac{1}{N} \sum_{i=1}^N \hat{g}_1(z_{1i}).$$

The estimated average treatment effect $\widehat{ATE}_{10}$ computed from marginal effects is asymptotically unbiased (Greene, 2007). Standard errors for the conditional and average treatment effects can be derived using the delta method (Raykov & Marcoulides, 2004). In Chapter 2, both the standard error and the delta method are described in more detail. Marginal effects are implemented in several statistical software packages, for example, the margins command in Stata (Williams, 2012), the margins package in R (Leeper, 2018), or within the Zelig approach (Imai, King, & Lau, 2008). The latter can be combined with the WhatIf package (Stoll, King, & Zeng, 2006) to investigate the model-dependency (as opposed to data-dependency) of the inference following a procedure of King and Zeng (2006). That is, if an outlier on the covariate produces an “extreme” expected count outcome – meaning: far away from the observed counts – inferences are said to be more model-dependent than when expected counts are similar to the observed counts.

Nevertheless, the marginal effects approach has two main shortcomings. First, the values of the covariate Z are assumed to be fixed-by-design. This implies, that the values of Z do not change from one sample to another, which is unlikely if individuals are randomly sampled. While the fixed-covariate assumption does not affect the estimation of regression coefficients in the Poisson regression model, the assumption may have an impact on the statistical inference on average and conditional treatment effects. This has previously been demonstrated for effect estimation in linear regression models (Chen, 2006; Kröhne, 2009) and will be further investigated in Chapter 2. Second, marginal effects can only be computed for manifest covariates, not for unobserved, latent variables. Latent variables are very common in psychology and other social sciences, where constructs-of-interest cannot

be directly observed, but are indirectly measured by multiple indicators (e.g., intelligence; measured with multiple items of a mental ability test). Controlling for latent variables instead of a fallible substitute (e.g., sum score) can be crucial for two reasons. First, measurement error leads to attenuation bias in regression coefficient estimates. Second, a causality condition that holds for a latent variable, does not necessarily hold for its fallible substitute (Sengewald et al., 2018). In these cases, only the latent covariate can be used to imply unbiasedness of the conditional effect function.

Other Regression-Based Methods

The g-computation approach by J. Robins (1986) is quite similar to marginal effects computation. While marginal effects are averages over conditional treatment effects, the g-computation regresses the unconditional expectation of the expected count outcomes on the treatment variable to obtain an average treatment effect. In formal notation, this would translate to

$$\frac{1}{N} \sum_{i=1}^N \hat{E}(Y|X=x, Z_1=z_{1i}) = \alpha_0 + \alpha_1 \cdot x,$$

where α_1 is the average treatment effect (Snowden, Rose, & Mortimer, 2011). G-computation, however, uncouples model estimation and treatment effect estimation in a way that expected outcomes of Y could also be estimated with nonparametric or machine learning techniques. Thus, $\hat{E}(Y|X=x, Z_1=z_{1i})$ is not necessarily estimated with a count regression model. For didactic introductions to g-computation, see Ahern, Hubbard, and Galea (2009) and Snowden et al. (2011). Standard errors for α_1 have to be obtained using bootstrap techniques (Daniel et al., 2011; Wang et al., 2017).

Instrumental variables are commonly used in the econometrics literature and their use for causal inference was described by Angrist, Imbens, and Rubin (1996) and Imbens (2014) to mention a few. Mullahy (1997) proposed an instrumental variable method for count regression models. In a nutshell, an instrumental variable is a covariate which (a) influences the choice or assignment of the treatment X , (b) has no direct effect on the outcome Y , but only indirectly through the treatment, and (c) has no common cause with the outcome variable Y (Lousdal, 2018). While the nature of instrumental variables exhibit the same intuition as covariates used for propensity scores (i.e., predicting treatment probability instead of the outcome), both approaches can yield discrepant results (Laborde-Cast  rot, Agrinier, & Thilly, 2015).

1.2.2 Propensity Score-Based Causal Inference

Propensity score-based methods usually aim to provide causal inferences by selecting covariates that satisfy the causality condition of a conditional independent cause, that is,

$$P(X=x|Z) = P(X=x|C_X).$$

The use of propensity scores for causal inferences has been described by [Rosenbaum and Rubin \(1983\)](#). For our simple example, with binary treatment variable X and unidimensional covariate Z_1 , a propensity score model can be formulated as logistic regression, namely,

$$P(X = 1|Z_1) = \frac{\exp(\beta_0 + \beta_1 Z_1)}{1 + \exp(\beta_0 + \beta_1 Z_1)}.$$

In this case, the probability of being treated depends on the value of the covariate Z_1 . To give an instance, if Z_1 was a pretest measure of depression and $X = 1$ a psychotherapy, the propensity score could reflect that persons with high pretest depression are more likely to self-select into psychotherapy. While it is common to use a logistic regression model to estimate propensity scores, the results can unduly be influenced by outliers in the covariates ([Pregibon, 1982](#)). An alternative are robit models ([J. H. Albert & Chib, 1993](#); [C. Liu, 2005](#)), which offer a more robust performance ([Kang & Schafer, 2006](#)). It is also possible to include latent variables as predictors in propensity score estimation ([Raykov, 2012](#)).

The count outcome variable Y does not appear in the propensity score equation. Consequently, for propensity score estimation it usually does not make much of a difference whether Y is continuous, count, or binary. Exceptions do exist as for example in propensity score matching, where the count nature of the outcome has to be addressed explicitly under some conditions.

Propensity Score Matching

Propensity score matching ([Rosenbaum, 2002](#); [Rubin & Thomas, 1996](#)) aims at selecting subsamples of treated and untreated persons with a similar covariate distribution. By way of illustration, if Z_1 was a pretest depression score, we would match treated and untreated persons with a similar probability of self-selecting into psychotherapy. For these matched subsamples, group comparisons can yield causally unbiased effect estimates. Matching algorithms are available in most statistical software packages, for example, in the Stata module PSMATCH2 ([Leuven & Sianesi, 2003](#)) and the R package MatchIt ([Ho, Imai, King, & Stuart, 2011](#)). Count regression models can be used to estimate treatment effects based on a matched subsample. In a Monte Carlo simulation, [Austin \(2009\)](#) found improved statistical inference of the exponentiated regression coefficients (i.e., rate ratio) approach in matched samples compared to unmatched samples. This effect, however, can be impaired if overdispersion is present in the count outcome ([Li, 2010](#)). [DeSantis, Lazaridis, Ji, and Spinale \(2013\)](#) introduced a zero-inflated Poisson regression model for cases in which the assumption of independent and identical distributed observations does not hold for the matched subsamples.

Inverse-Propensity Weighting

Inverse-propensity weighting accounts for unequal probabilities of treatment by reweighting the participants. In this method, the inverse of the propensity score variable is used as sampling weight (Hirano & Imbens, 2002; Schafer & Kang, 2008). The intuition behind this procedure is to balance the covariates' distribution by de-emphasizing observations that are very likely to be treated, and emphasizing observations with low probability for treatment. The inverse-propensity weighting can be biased, if estimated propensity scores are close to 0 or 1, because the weighting procedure is sensitive for such outliers (Schafer & Kang, 2008). In addition, even if the inverse-propensity weighting results in causally unbiased effect estimates, it is often statistically inefficient (Ertefaie & Stephens, 2010). If inverse-propensity weights are incorporated in a statistical analysis, the standard errors should reflect that these weights are estimated. For linear models, J. M. Robins, Rotnitzky, and Zhao (1995) proposed a method to compute standard errors accounting for estimated weights.

1.2.3 Double Robust Methods

In recent years, several approaches combining regression- and propensity score-based techniques have been suggested. They are often advertised with a property named *double robustness* because it suffices if either the regression model or the propensity score model is correct to obtain causally unbiased treatment effect estimates. These double robust methods aim to satisfy two causality conditions, but it is sufficient if only one condition holds. They can be useful to achieve bias reduction *and* improved efficiency in the effect estimation. Any combination of regression- and propensity score-based techniques can be used to construct a double robust method and we will not discuss them here in detail. For an overview of construction and performance of double robust methods, see Kang and Schafer (2006). Well-known approaches in this category include targeted maximum likelihood estimation originally proposed by van der Laan and Rubin (2006) and augmented inverse-propensity weighting originally proposed by J. Robins, Rotnitzky, and Zhao (1994). For accessible introductions to these techniques, we refer to Glynn and Quinn (2010) and Schuler and Rose (2016).

1.3 Statistical Frameworks for Count Outcomes

In the previous section, we introduced a parameterization for the conditional expectation $E(Y|X, Z)$ and the corresponding conditional effect function $g_t(Z)$. In this section, we give an overview of statistical frameworks to estimate count regression models. For simplicity, we continue our simple example with a binary treatment X and a covariate Z_1 from Equation (1.4) which was given by:

$$E(Y|X, Z_1) = \exp(\gamma_{00} + \gamma_{10}X + \gamma_{01}Z_1 + \gamma_{11}XZ_1).$$

1.3.1 Family of Poisson Regression Models

The original Poisson regression model was introduced by McCullagh and Nelder (1996) as part of the family of generalized linear models. In the notation of McCullagh and Nelder (1996), the Poisson regression model for a sample size N is formulated:

$$\begin{aligned} E(Y_i) &= \mu_i \\ \eta_i &= \mathbf{X}_i \boldsymbol{\gamma} \\ \log(\mu_i) &= \eta_i, \end{aligned}$$

where Y_i is the i th observations of Y ; μ_i is the exponential transformation of the linear predictor η_i ; the linear predictor η_i is given by $\mathbf{X}_i \boldsymbol{\gamma}$; $\mathbf{X}_i$ is the i th row of the design matrix $\mathbf{X}$. Further, the outcome is assumed to be Poisson distributed $Y_i \sim \mathcal{P}(\lambda = \mu_i)$. For a gentle introduction to Poisson regression models, see Cox et al. (2009) and Dobson (2001).

Our simple example with count outcome Y , binary treatment variable X , and one-dimensional Z_1 from Equation (1.4) can be represented as Poisson regression model as follows:

$$\begin{pmatrix} \log[E(Y_1)] \\ \log[E(Y_2)] \\ \vdots \\ \log[E(Y_N)] \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & z_{11} & x_1 z_{11} \\ 1 & x_2 & z_{12} & x_2 z_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & z_{1N} & x_N z_{1N} \end{pmatrix} \cdot \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{01} \\ \gamma_{11} \end{pmatrix}, \quad (1.5)$$

where x_i and z_{1i} are observations of the variables X and Z_1 . In the Poisson regression model, the treatment variable, the covariate, and their interaction are treated as fixed-by-design instead of as random variables. This *fixed design matrix assumption* spares us the need to specify distributional assumptions on the covariate when estimating the regression coefficients. Nevertheless, the maximum likelihood estimation of the regression parameters $\boldsymbol{\gamma}$ given by the log-likelihood

$$\log L(\boldsymbol{\gamma} | Y, X, Z_1) = \sum_{i=1}^N y_i \cdot \log(\eta_i) - \log(y_i!) - \eta_i$$

and their standard errors remain unaffected by this assumption (for a proof, see Appendix 2.B of Chapter 2).

In recent years, the *family of Poisson regression models* has been extended by several means. The Poisson distribution has been substituted by other count distributions, for example, the negative binomial (NB) distribution (i.e., negative binomial regression). In contrast to the Poisson distribution, where $E(Y_i) = \text{Var}(Y_i)$ (i.e., *equidispersion*), the NB distribution allows for various alternative variance parameterizations (for an overview, see Hilbe, 2011, p. 284). Maybe the best known variance parameterization is the NB2 model,

where $Var(Y_i) = \mu_i + \alpha\mu_i^2$ for $\alpha \geq 0$ (i.e., *overdispersion*; the variance is greater than the expectation). For $\alpha = 0$, the NB2 model is equivalent to a standard Poisson regression model.

1.3.2 Semiparametric Count Regression and M-Estimation

Our simple count regression model from Equation (1.4) can also be estimated using a semiparametric approach. In contrast to the family of Poisson regression models, a semiparametric count regression model is not confined to a single conditional outcome distribution (e.g., Poisson distribution), but enables a (possibly infinite-dimensional) space of count distributions. A semiparametric estimator for our example can be obtained by solving the vector of estimating equations

$$\begin{aligned} E[\psi(Y, X, Z_1, \gamma)] &= E[\mathbf{D}(X, Z_1, \gamma) [Y - \exp(\gamma_{00} + \gamma_{10}X + \gamma_{01}Z_1 + \gamma_{11}XZ_1)]] \quad (1.6) \\ &= \mathbf{0}, \end{aligned}$$

where $\mathbf{D}(X, Z_1, \gamma)$ is an arbitrary 4×1 vector of functions of the treatment variable X , the covariate Z_1 , and the parameters γ (Tsiatis, 2006). These kinds of estimators are also known as M-estimators (Stefanski & Boos, 2002), which also include the class of maximum likelihood estimators. For example, if we take $\mathbf{D}(X, Z_1, \gamma) = (1, X, Z_1, XZ_1)^T$, the solution of the estimating equations in (1.6) is equivalent to the estimated parameters from a Poisson regression model.

This approach has two advantages. First, the so-called sandwich estimator of the asymptotic variance of estimated parameters is robust against misspecifications of the model and is given by

$$\mathbf{V}(\gamma) = \mathbf{A}(\gamma)^{-1} \mathbf{B}(\gamma) (\mathbf{A}(\gamma)^{-1})^T,$$

where $\mathbf{A}(\gamma) = E[-\psi'(Y, X, Z_1, \gamma)]$ and $\mathbf{B}(\gamma) = E[\psi(Y, X, Z_1, \gamma)\psi(Y, X, Z_1, \gamma)^T]$. For example, count data often exhibit overdispersion which would lead to biased standard error estimates in a Poisson regression model (assuming equidispersion). If we take $\mathbf{D}(X, Z_1, \gamma) = (1, X, Z_1, XZ_1)$, the point estimates of the semiparametric approach would be equivalent to a Poisson regression model, yet the standard errors would account for the overdispersion. Second, the estimating equations enable us to include additional parameters of interest or transformations of parameters in the model estimation. In Chapter 2, for example, we will add estimating equations for group size parameters θ to the count regression model, and thus obtain a joint (robust) covariance matrix $\mathbf{V}(\gamma, \theta)$. Further variations of semiparametric count regression models have been proposed, for example, by Basu and Rathouz (2005) and Gurmu, Rilstone, and Stern (1999). For further information on semiparametric modeling, see Tsiatis (2006).

1.3.3 Structural Equation Modeling

The extensions of the moment-based approach in Chapters 3 and 4 are based on structural equation modeling (SEM; Bentler, 1980; Bollen, 1989), a multivariate statistical framework with two central parts:

$$\boldsymbol{\eta} = \boldsymbol{\beta}_0 + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta} \quad (1.7)$$

$$\mathbf{z} = \boldsymbol{\nu} + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (1.8)$$

In Equation (1.7), the structural part of the model is presented, which describes the (linear) relationships among the variables $\boldsymbol{\eta}$ with intercept vector $\boldsymbol{\beta}_0$, matrix of regression coefficients $\mathbf{B}$, and vector of regression residuals $\boldsymbol{\zeta}$. In contrast to ordinary least square regression models, SEM allows for multiple simultaneous linear regressions among the variables $\boldsymbol{\eta}$. In Equation (1.8), the measurement part of the model is presented. The variables $\mathbf{z}$ are fallible measurements of the common factors or latent variables $\boldsymbol{\eta}$ with intercept vector $\boldsymbol{\nu}$ and matrix of factor loadings $\mathbf{\Lambda}$. The measurement model enables a decomposition of the observed values of $\mathbf{z}$ into the unobserved *true* values of the latent variable $\boldsymbol{\eta}$ and the measurement error $\boldsymbol{\varepsilon}$. Consequently, the regressions of the structural model are purged from measurement error. The observed variables $\mathbf{z}$ are usually assumed to follow a multivariate normal distribution and hence the log-likelihood for a basic SEM is

$$\log L(\boldsymbol{\theta}|\mathbf{z}) = -\frac{NL}{2} \log(2\pi) - \frac{N}{2} \log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) - \frac{1}{2} \sum_{i=1}^N (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta}))^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} (\mathbf{z}_i - \boldsymbol{\mu}(\boldsymbol{\theta})),$$

where $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ are the model-implied mean and covariance structure, and L is the dimension of $\mathbf{z}$. In Chapters 3 and 4, we will show in detail how structural equation models can be extended for count regression models and multigroup analysis.

1.3.4 Bayesian Count Regression Models

So far and throughout the thesis, we confined ourselves to so-called frequentist methods of statistical inference. An alternative are Bayesian statistics, which adopt a different perspective on statistical modeling. For example, while the aforementioned statistical frameworks pose the existence of a true and fixed parameter $\boldsymbol{\gamma}$, these parameters are random variables within the Bayesian framework. As a consequence, Bayesian procedures include a prior distribution (sometimes called prior information or belief) of the parameters, and their posterior distribution given the data is estimated.

Bayesian statistics are not just another framework for estimating parameters, but a different mindset of statistical thinking, which we will not delve into in this thesis. There are Bayesian variants of traditional Poisson regression models (El-Sayyad, 1973), Bayesian negative binomial regression models (Hilbe, 2011, Chapter 15), Bayesian semiparametric count regression models (Carota, 2002; Dunson, 2005), and Bayesian structural equation

modeling (Lee, 2007; Muthén & Asparouhov, 2012). We will discuss Bayesian SEM as alternative statistical framework for the moment-based approach in Chapter 5. For an overview of Bayesian statistics, see Gelman, Carlin, Stern, and Rubin (2004).

1.4 Goals and Structure of this Thesis

The main developments and extensions of a new approach for estimating average and conditional treatment effects on count outcomes are motivated by the key shortcomings of the traditional approaches, in particular exponentiated regression coefficients and marginal effects, presented in the previous sections: First, ambiguous effect definitions using ratios of conditional expectations. Second, neglect of sampling variability of the covariates and thus, impaired statistical inferences. Third, exclusion of latent variables and, thereby, not accounting for measurement error in the covariates. Consequently, the goal of this thesis is to develop and provide a comprehensive approach for estimating average and conditional treatment effects on count outcomes, which offers unambiguous effect definitions, improved statistical inference by accounting for sampling variability in the covariates, and allows for latent covariates. We pursued this goal with a step-wise development and enhancement of the so-called *moment-based approach*. The thesis is structured as follows:

In Chapter 2, the moment-based approach is introduced for our simple example with count outcome Y , binary treatment variable X , and one-dimensional covariate Z . In contrast to the marginal effects approach, the average treatment effect is not computed as sample average of estimated conditional effects, but directly based on sample moments of the covariate. The analytical formulas required for this technique are derived from the moment-generating function of the covariate Z . Therefore, a distributional assumption for the covariate Z has to be made. In a simulation study, performances of the moment-based approach and the marginal effects approach are compared.

In Chapter 3, the moment-based approach is extended to scenarios with a multidimensional, latent covariate η . The statistical framework changes from a generalized linear model to a negative binomial multigroup structural equation model. Within this framework, the estimation of negative binomial regressions with latent covariates is possible. To derive analytical effect formulas, the covariates are assumed to be (conditional) multivariate normally distributed. In an illustrative example, the extended moment-based approach is applied to investigate effectiveness of a cognitive training for elderly persons.

In Chapter 4, the assumption of (conditional) multivariate normally distributed covariates is relaxed. We present four different alternatives to account for non-normally distributed covariates. Similar to the EffectLiteR approach (Mayer et al., 2016), we distinguish between categorical covariates K and continuous covariates Z . For non-normal continuous covariates, we suggest three solutions depending on how much information on their joint distribution is available. While in the previous chapters we focused on the average treatment effect of a binary treatment variable, we now consider $p + 1$ levels of the treatment variable X and compute aggregated effects besides the average treatment

effect (e.g., conditional effect given a gender).

In Chapter 5, a general discussion of the developments in this thesis is given. We summarize the extensions and features of the moment-based approach and provide guidelines in which scenarios its application can be beneficial. Additionally, we will discuss limitations of the proposed moment-based approach and have an outlook on possible further developments. Alternative estimation techniques beside classical maximum likelihood estimation as well as potential generalizations of the moment-based approach to regression models with other link functions (e.g., logistic) are considered.

2 A New Approach with Stochastic Covariates

Abstract

Researchers often use regressions with a logarithmic link function to evaluate the effects of a treatment on a count variable. In order to judge the average effectiveness of the treatment on the original count scale, they compute average treatment effects, which are defined as the average difference between the expected outcomes under treatment and under control. Current practice is to evaluate the expected differences at every observation and use the sample mean of these differences as a point estimate of the average effect. The standard error for this average effect estimate is based on the implicit assumption that covariate values are fixed, i.e., do not vary across different samples. In this paper, we present a new way of analytically computing average effects based on regressions with log link using stochastic covariates and develop new formulas to obtain standard errors for the average effect. In a simulation study, we evaluate the statistical performance of our new estimator and compare it with the traditional approach. Our findings suggest that the new approach gives unbiased effect estimates and standard errors and outperforms the traditional approach when strong interaction and/or a skewed covariate is present.

2.1 Introduction

The evaluation of treatment effects on count variables is quite common in psychology. For example, clinical psychologists have examined treatment effects on the counts of classroom violations and instances of reinforcing peers' negative behavior in children with ADHD (Helseth et al., 2015; Pelham, Jr. et al., 2016) as well as instances of binge eating among patients with eating disorders (Kenny et al., 2018). In the field of drug and substance use prevention, the effects of interventions on the weekly sum of alcoholic drinks and cigarettes consumed have been examined (Layland et al., 2018; Muench et al., 2017). Organizational psychologists have examined days of absenteeism (Johns & Al Hajj, 2016; Schaumberg & Flynn, 2017). Further examples are counts of wrongly answered items on a test (Macdonald et al., 2017), counts of traffic violations (Precht et al., 2017), and counts of risky sexual behavior (Hittner, Owens, & Swickert, 2016; Schrimshaw, Antebi-Gruszka, & Downing, 2016).

Analyzing the effects of a treatment X on a count variable Y can be challenging, especially if covariates Z and treatment by covariate interactions XZ are also considered. It is well known by now that traditional approaches like ordinary least square (OLS) regressions which treat the dependent variable as continuous are unsuitable for two reasons. First, OLS regression is likely to predict negative conditional expectations of the dependent variable for some values of the covariates. As counts are limited to nonnegative integers, these negative predicted values cannot be interpreted meaningfully. Second, the assumptions regarding residuals in OLS regression are likely to be violated when the dependent variable is a count variable. Count data residuals are discrete and often highly skewed and heteroscedastic (Coxe et al., 2009; Gardner, Mulvey, & Shaw, 1995), while OLS regression requires normally distributed, homoscedastic residuals. Thus, OLS regression might yield less meaningful predictions and unreliable inferences for count data.

The generalized linear model (GLM) with logarithmic link function and negative binomial random component has been embraced as an appropriate model for count outcomes (Coxe et al., 2009; Gardner et al., 1995; Hilbe, 2007). It consists of a linear function of regressors (systematic component) which is linked to the logarithm of the conditional expectation of the count outcome. The counts are assumed to follow a conditional negative binomial distribution, and the regressors in the design matrix are assumed to be fixed, i.e., the values of the regressors in the design matrix do not vary over repeated sampling. Negative binomial (NB) GLM can account for both the fixed lower bound of count outcomes and their skewed, heteroscedastic residuals and, therefore, provides appropriate estimates of regression coefficients. The well-known Poisson regression is a special case of NB-GLM.

It is important to note that the coefficients of regressions with log link, like those of any nonlinear regression model, do not correspond directly to (average) treatment effects on the original count scale of the outcome (Greene, 2007, p. 729). Hence, treatment effects are often computed in a subsequent step by comparing conditional expectations of the outcome under control and under treatment (Steyer & Nagel, 2017, p. 415). Researchers

can either consider effects based on ratios of conditional expectations or effects based on differences in conditional expectations. A commonly used strategy is to exponentiate the regression coefficient of the treatment variable X , which then corresponds to the *ratio* of expected counts under treatment and under control (Agresti, 2007, p. 79). Treatment effects based on ratios (e.g., risk or rate ratio) are non-collapsible for regressions with log link (Gail, Wieand, & Piantadosi, 1984). Otherwise stated, the average of conditional (or individual) effects does not always equal the ratio of group averages. Thus, in randomized controlled treatments, the comparison of group averages does not necessarily reflect the average treatment effect. We will discuss this in more detail in the next section and in “Appendix 2.A”.

An alternative is to consider the effects of binary independent variables based on differences, i.e., effects that are defined as the *difference* between the expected counts under treatment and under control. These effects are collapsible (i.e., average over individual differences equals the difference of averages), and their interpretation is easier. Therefore, we focus on effects based on differences in this paper. In the literature, average effects based on differences are sometimes also referred to as (average) marginal effects or average partial effects (Long, 1997). Current practice for estimating such effects is to evaluate the expected difference at every observation based on the NB-GLM and use the sample mean of these differences as the point estimate of the average effect (Greene, 2007, p. 730). Again, this is a two-step procedure in which first, a regression model is estimated and, second, the differences between the estimated conditional expectations are computed. Both steps are typically based on the implicit assumption that covariate values are fixed, i.e., do not vary across samples.

The fixed-covariate assumption is often unrealistic – especially for continuous covariates – but has received little attention in the literature because linear models with fixed regressors, for instance, yield identical point estimates and standard errors for regression coefficients as models with stochastic regressors (Renchner & Schaalje, 2007; Sampson, 1974). However, it has been shown that the fixed-covariate assumption can make a difference, for example, when evaluating the predictive validity of a model (Rosset & Tibshirani, 2018), in power calculations (Gatsonis & Sampson, 1989), or when computing simple slopes (Y. Liu et al., 2017). Thus, standard errors of average effects are likely to be underestimated when covariates are in fact stochastic, i.e., when sampling variability in covariate values is not taken into account. While the magnitude of the underestimation of standard errors due to fixed covariates has been examined for linear regression models (Chen, 2006; Crager, 1987; Kröhne, 2009; Nagengast, 2006; Sampson, 1974), little is known about potential underestimation in nonlinear models.

In this paper, we present a new approach to estimating the average treatment effect for regressions with log link and a single stochastic covariate. This approach is based on an analytical computation of the average treatment effect and allows us to develop new formulas for obtaining point estimates and standard errors for the average effect. In contrast to traditional approaches, we no longer assume fixed values for covariates but

need to make a distributional assumption about the covariates. Furthermore, we use Monte Carlo simulations to evaluate the statistical performance of the new approach and the accuracy of its standard errors when covariates are in fact stochastic. Our aim is to study whether (and to what extent) the fixed-covariate assumption of the traditional ATE estimator leads to underestimated standard errors for average effects and whether the new estimator yields more accurate standard errors.

2.1.1 Definition and Terminology of Effects

Throughout this paper, we focus on effects based on differences between (conditional) expectations. We begin by introducing our definitions and terminology with respect to a case with a single covariate. We distinguish between three (potentially) different effects that can be of interest for applied researchers: (1) the unconditional treatment effect (UTE), (2) the average treatment effect (ATE), and (3) the treatment effect at the mean of Z (TEMZ). In a simple example with a binary treatment variable X ($X=0$: control group; $X=1$: treatment group) and a single continuous covariate Z with $E(Z) = \bar{Z}$, these effects are defined as follows:

$$\begin{aligned} \text{UTE} &= E(Y|X=1) - E(Y|X=0), \\ \text{ATE} &= E[E(Y|X=1, Z) - E(Y|X=0, Z)] \end{aligned} \tag{2.1}$$

$$= E[E(Y|X=1, Z)] - E[E(Y|X=0, Z)], \tag{2.2}$$

$$\text{TEMZ} = E(Y|X=1, Z=\bar{Z}) - E(Y|X=0, Z=\bar{Z}).$$

In randomized controlled trials, where treatment assignment is stochastically independent of all pre-treatment covariates, we do not need to control for covariates and can consider the unconditional treatment effect (UTE). However, it can sometimes be beneficial to include additional covariates even in randomized experiments, for example, to increase power or when randomized treatment assignment is believed to have been compromised by systematic dropout or other factors. The inclusion and careful selection of covariates becomes indispensable in observational studies where treatment assignment depends on individuals' characteristics. In models with one or multiple covariates, we can either focus on the average of the conditional treatment effects (ATE) or look at the treatment effect at the mean of Z (TEMZ). In the general case for models with nonlinear link functions, the three effects can be different, but there are two noteworthy special cases: In randomized controlled trials, the UTE is identical to ATE (for more details on causality conditions see [Mayer, Thoemmes, et al., 2014](#)) and in models with linear link functions, the ATE is identical to TEMZ.

For effects defined as ratios, the analogues to Eqs. (2.1) and (2.2) are no longer equal, i.e.,

$$E \left[\frac{E(Y|X=1, Z)}{E(Y|X=0, Z)} \right] \neq \frac{E[E(Y|X=1, Z)]}{E[E(Y|X=0, Z)]}$$

except for two special cases when either there is no covariate effect in the control group or no interaction effect. We provide a proof for this in “Appendix 2.A”. As a consequence, the average treatment effect defined on the basis of individual treatment effects does not necessarily equal the ratio of group averages. In the literature, this phenomenon is often discussed under the term non-collapsibility. Nevertheless, for effects based on differences [i.e., Eqs. (2.1) and (2.2)], the difference of Z -adjusted group means is always equal to the mean over all Z -conditional differences.

2.2 Regression with Log Link

We use the term *regression with a logarithmic link function* (log link) for a regression in which the regressors are related to the logarithm of the conditional expectation of the count outcome. A regression with log link can be expressed by two equivalent equations

$$\begin{aligned}\log[E(Y|Z_1, \dots, Z_k)] &= \beta_0 + \beta_1 Z_1 + \dots + \beta_k Z_k \\ E(Y|Z_1, \dots, Z_k) &= \exp(\beta_0 + \beta_1 Z_1 + \dots + \beta_k Z_k),\end{aligned}$$

where Y is a nonnegative random variable and $Z_i, i = 1, \dots, k$ are predictor variables. The random variable Z_i can also be a function of other predictors, that is, interaction or polynomial terms can be included as well. Note that the derivations in this section apply to regressions with log link in general, and not to a specific regression model with log link (e.g., NB-GLM).

For a regression with log link with only a dichotomous treatment variable $Z_1 = X$ and a count outcome Y , the regression equation can be written as:

$$E(Y|X) = \exp(\beta_0 + \beta_1 X). \quad (2.3)$$

Adding a unidimensional continuous covariate $Z_2 = Z$ and including an interaction term $Z_3 = XZ$, we will also consider conditional regressions with log link of the form:

$$E(Y|X, Z) = \exp(\beta_{00} + \beta_{10}X + \beta_{01}Z + \beta_{11}XZ) \quad (2.4)$$

2.2.1 Treatment Effects in Regressions with Log Link

For the aforementioned regressions with log link, our three effects of interest are given by (based on Eq. (2.4))¹:

$$\text{UTE} = E[\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z)|X = 1] - E[\exp(\beta_{00} + \beta_{01}Z)|X = 0] \quad (2.5)$$

$$\text{ATE} = E[\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z)] - E[\exp(\beta_{00} + \beta_{01}Z)] \quad (2.6)$$

$$\text{TEMZ} = \exp(\beta_{00} + \beta_{10} + \beta_{01}\bar{Z} + \beta_{11}\bar{Z}) - \exp(\beta_{00} + \beta_{01}\bar{Z}). \quad (2.7)$$

¹The UTE can also be expressed as in Eq. (2.3): $\text{UTE} = \exp(\beta_0 + \beta_1) - \exp(\beta_0)$. This notation is always equivalent to Eq. (2.5). For more information, see Mayer, Thoemmes, et al. (2014).

Again, UTE and ATE are equivalent for randomized control trials. The UTE is computed over the $X = 0$ - and $X = 1$ -conditional distributions of Z , while the ATE is computed over the unconditional distribution of Z . For randomized control trials, these three distributions are the same. But ATE and TEMZ are not equal for regressions with log link in general, because Jensen's inequality (Jensen, 1906) applies. That is,

$$E[\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z)] \geq \exp(\beta_{00} + \beta_{10} + \beta_{01}\bar{Z} + \beta_{11}\bar{Z}) \quad (2.8)$$

$$E[\exp(\beta_{00} + \beta_{01}Z)] \geq \exp(\beta_{00} + \beta_{01}\bar{Z}), \quad (2.9)$$

all of the components of the TEMZ are smaller than the corresponding components of the ATE. Thus, unlike in linear regression models the TEMZ is not an appropriate substitute for the ATE. In other words, the treatment effect at the mean of Z is not equal to the average of the $Z = z$ -conditional treatment effects. To illustrate the idea of averaging the $Z = z$ -conditional treatment effects, we can rewrite the ATE as an integral:

$$\begin{aligned} \text{ATE} &= E[E(Y|X=1, Z) - E(Y|X=0, Z)] \\ &= \int (\exp[\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z] - \exp[\beta_{00} + \beta_{01}Z]) P_Z(dz), \end{aligned} \quad (2.10)$$

i.e., an integral (or sum) of conditional effects weighted by their density. The probability of a conditional effect is determined by the distribution of the covariate Z . Hence, the distribution of Z is needed to obtain this expectation (and not just the mean, as is the case for the TEMZ).

2.3 Estimation of Regression Coefficients

In practice, the regression coefficients of regressions with log link are unknown and need to be estimated from a sample. Let $\mathbf{y} = (Y_1, Y_2, \dots, Y_N)^T$, $\mathbf{x} = (X_1, X_2, \dots, X_N)^T$, $\mathbf{z} = (Z_1, Z_2, \dots, Z_N)^T$ represent a sample of N independent observations. For estimation, we use a parametric statistical model and denote the vector of regression coefficients by $\boldsymbol{\beta} = (\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11})^T$.

Generalized linear models are widely used to estimate $\boldsymbol{\beta}$. These models require the specification of a conditional distribution of Y_i . Common choices for analyzing count data include the Poisson distribution $\mathcal{P}(\lambda)$ (Poisson GLM) and the negative binomial distribution $\text{NB}(r, p)$ (negative binomial (NB) GLM; for more information, see Hilbe, 2007). The latter distribution is generally more flexible, because it can deal with overdispersion, i.e., cases where the variance exceeds the mean. This is particularly relevant when there is unobserved heterogeneity that is not fully accounted for by the observed covariates.

The maximum likelihood estimates of regression coefficients $\hat{\boldsymbol{\beta}}$ are usually only based on the conditional distribution of Y_i , but not on the joint distribution of X_i and Z_i (Hilbe, 2007). Consequently, the likelihood of the regressors does not affect the estimation of $\boldsymbol{\beta}$. The estimation of the regression coefficients is asymptotically unbiased when the

assumptions of the model hold. However, the estimates may be biased in finite samples, especially in small samples (e.g., $N = 50$; Lord, 2006).

Other models have been suggested with different conditional distributions of Y_i , for example, overdispersed Poisson regression, Gamma regression (Ng & Cribbie, 2016), or semi-parametric approaches (Basu & Rathouz, 2005). Carefully deciding on an appropriate regression model is crucial in order to obtain unbiased regression coefficients estimates, which are needed for effect estimation.

2.4 Estimates and Standard Errors of Treatment Effects

In regressions with log link, the regression coefficients do not directly correspond to the average effect of interest. To estimate an ATE, we not only need estimates of the regression coefficients, but also have to consider the (unconditional) distribution of Z . Generally, there are two approaches to account for the unconditional distribution of Z : first, using the empirical distribution of the observed z_i or, second, making an assumption about the unconditional distribution of Z . While traditional ATE estimators rely on the former, our new approach is based on the latter.

2.4.1 Empirical-Distribution-Based Approach

The traditional approach to estimating the ATE relies on a numerical approximation of Eq. (2.10):

$$\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^N \exp(\hat{\beta}_{00} + \hat{\beta}_{10} + \hat{\beta}_{01}z_i + \hat{\beta}_{11}z_i) - \exp(\hat{\beta}_{00} + \hat{\beta}_{01}z_i). \quad (2.11)$$

Here, the ATE is approximated by a finite sum of values weighted by the empirical distribution of the covariates. This is a consistent estimator of the ATE, that is, the estimation will be unbiased with increasing sample size (Greene, 2007). We call this the empirical approach.

The function $\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta})$ only depends on the regression coefficients $\hat{\beta}$. The z_i are observed values of Z_i and treated as constants that have to be inserted, that is, they are not considered as random variables or estimates. Note that this fixed-covariate assumption does not necessarily depend on whether the covariates were treated as fixed or random in the estimation of $\hat{\beta}$. Thus, even if $\hat{\beta}$ was estimated assuming stochastic covariates, $\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta})$ would still implicitly assume fixed covariates.

The empirical approach is illustrated in Fig. 2.1b with a sample of size $N = 10$. The individual expectations $E(Y_i)$ with either $x_i = 1$ (i.e., the individual expectation under treatment) or $x_i = 0$ (i.e., the individual expectation under control) are given, resembling the two parts of Eq. (2.11). For each person i in the sample the covariate value z_i is used to compute the individual difference between the two predicted outcomes. These estimated conditional effects $\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta})$ are given in Table 2.1. The ATE is computed

Table 2.1: Illustrative example of how the ATE is computed using the empirical approach.

i	z_i	$\hat{E}(Y X = 1, Z = z_i)$	$\hat{E}(Y X = 0, Z = z_i)$	$\hat{\delta}_i$
1	0.252	1.103	0.843	− 0.259
2	1.130	1.550	1.630	0.079
3	1.057	1.507	1.543	0.036
4	0.739	1.332	1.216	− 0.116
5	1.690	1.927	2.481	0.555
6	− 0.059	0.977	0.668	− 0.309
7	1.816	2.023	2.726	0.703
8	0.787	1.357	1.260	− 0.097
9	1.720	1.949	2.537	0.588
10	0.868	1.400	1.339	− 0.061
				$\widehat{\text{ATE}}_{\text{EMP}} = \frac{1}{10} \sum_{i=1}^{10} \hat{\delta}_i = 0.112$

Estimated regression coefficients used in $\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta})$ are $\hat{\beta}_{00} = 0$, $\hat{\beta}_{01} = 0.388$, $\hat{\beta}_{10} = -0.359$, and $\hat{\beta}_{11} = 0.362$. $\hat{\delta}_i$ denotes the difference in conditional expectations, i.e., $\hat{\delta}_i = \hat{E}(Y|X = 1, Z = z_i) - \hat{E}(Y|X = 0, Z = z_i)$

by averaging the conditional effects. In contrast, the TEMZ can be thought of as the conditional effect at the mean of $\bar{z}$ (illustrated in Fig. 2.1a). In this case, the TEMZ is substantively smaller than the ATE and indicates that the treatment had no average effect.

The squared standard error of $\widehat{\text{ATE}}_{\text{EMP}}(\hat{\beta})$ is given by

$$\text{SE}_{\text{EMP}}^2 = \left(\frac{\partial \widehat{\text{ATE}}_{\text{EMP}}}{\partial \hat{\beta}} \right)^T \mathbf{V}(\hat{\beta}) \frac{\partial \widehat{\text{ATE}}_{\text{EMP}}}{\partial \hat{\beta}},$$

where $\mathbf{V}(\hat{\beta})$ is the covariance matrix of regression parameter estimates (Dowd, Greene, & Norton, 2014). It can be obtained, for example, from the NB-GLM model. As the z_i are considered fixed, and not estimated, their potential variation across samples is neglected in SE_{EMP}^2 .

2.4.2 Moment-Based Approach with Stochastic Covariates

We suggest a moment-based estimation of the ATE that allows the regressors X and Z to be modeled as stochastic. The new approach relies on an assumption about the distribution of Z and can be derived from Eq. (2.6):

$$\begin{aligned} \text{ATE} &= E[\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z)] - E[\exp(\beta_{00} + \beta_{01}Z)] \\ &= \exp(\beta_{00} + \beta_{10})E[\exp((\beta_{01} + \beta_{11})Z)] - \exp(\beta_{00})E[\exp(\beta_{01}Z)] \\ &= \exp(\beta_{00} + \beta_{10})M_Z(\beta_{01} + \beta_{11}) - \exp(\beta_{00})M_Z(\beta_{01}). \end{aligned} \quad (2.12)$$

This representation allows us to compute the average effect using a function $M_Z(s)$, $s \in \mathbb{R}$. The function $M_Z(s)$ is called the moment-generating function (MGF) of the random vari-

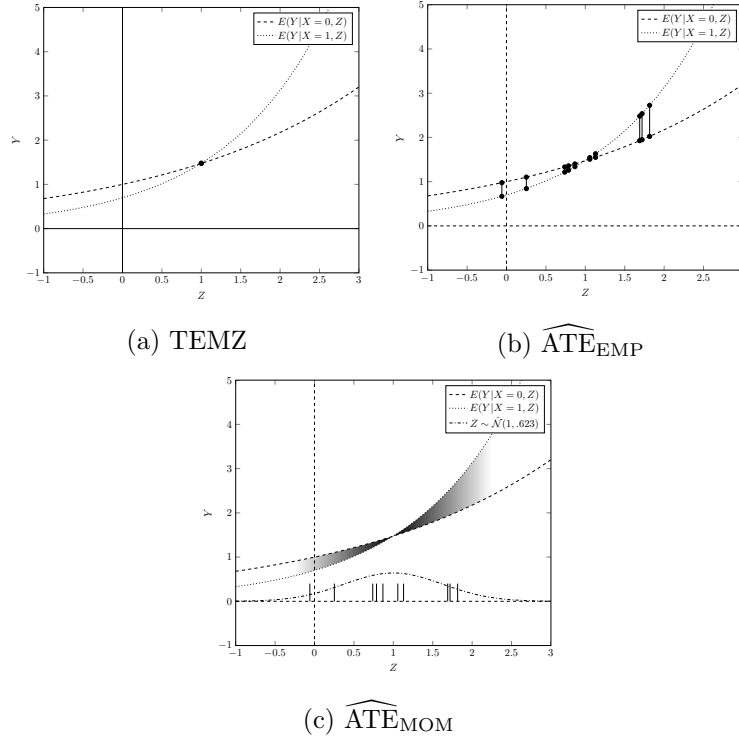


Figure 2.1

Plots for our illustrative example showing the group-specific regressions of Y on Z (dotted and dashed line) as well as the Z -conditional effects of the treatment X (lines with dot marks). **a** Only the conditional effect at the mean of z_i , that is, $\bar{z} = 1$ is inspected to compute the TEMZ. In this case, $TEMZ = 0$. **b** Conditional effects for all observed values z_i are shown. The average over these ten differences is the $\widehat{ATE}_{EMP} = 0.112$. **c** Covariate values (indicated by rug plot) are used to estimate parameters of normal distribution (dash dotted line). The shaded area between the regression lines represents the integral over conditional effects weighted by the density of Z .

Table 2.2: Moment-generating functions for several unconditional distributions of Z .

Distribution of Z	Moment-generating function
Binomial $\mathcal{B}(p)$	$M_Z(s) = 1 - p + pe^s$
Chi-square χ_n^2	$M_Z(s) = \frac{1}{(1-2s)^{n/2}}$
Gamma $\gamma(p, b)$	$M_Z(s) = \left(\frac{b}{b-s}\right)^p$
Uniform $U[a, b]$	$M_Z(s) = \frac{e^{sb} - e^{sa}}{s(b-a)}$
Normal $\mathcal{N}(\mu, \sigma^2)$	$M_Z(s) = \exp\left(\mu s + \frac{\sigma^2 s^2}{2}\right)$
Poisson $\mathcal{P}(\lambda)$	$M_Z(s) = \exp(\lambda(e^s - 1))$
Negative binomial $\text{NB}(r, p)$	$M_Z(s) = \left(\frac{pe^s}{1-(1-p)e^s}\right)^r$

Z is a continuous real-valued random variable, $s \in \mathbb{R}$ and the parameters of the distributions are $p \in [0, 1]$ for binomial distribution, $n \in \mathbb{N}$ for χ^2 -distribution, $p, b \in \mathbb{R}^+$ for gamma distribution, $a, b \in \mathbb{R}, a < b$ for uniform distribution, $\mu, \sigma^2 \in \mathbb{R}, \sigma^2 > 0$ for normal distribution, and $\lambda \in \mathbb{R}^+$ for Poisson distribution

able Z and is defined as $M_Z(s) = E[\exp(sZ)]$. Its derivatives are usually used to compute the moments of Z . However, (most) moment-generating functions have an explicit, closed form depending on the distribution of Z , which allows the ATE to be computed directly (without using an integral like in Eq. (2.10) or an approximation like the empirical approach). Table 2.2 gives an overview of exemplary univariate distributions of Z and their moment-generating functions.

Using moment-generating functions as in Eq. (2.12) allows us to formulate an estimator that relies both on estimated regression coefficients $\hat{\beta}$ and on p estimated parameters of the distribution $\hat{\theta}_i$, $i = 1, \dots, p$ of Z . The result is that we obtain $\widehat{M}_Z(s)$ as a moment-generating function with estimated parameters. In conclusion, the new ATE estimator can be written as:

$$\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) = \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \widehat{M}_Z(\hat{\beta}_{01} + \hat{\beta}_{11}) - \exp(\hat{\beta}_{00}) \widehat{M}_Z(\hat{\beta}_{01})$$

In contrast to the empirical ATE estimator, our approach does not rely on the manifest values z_i , but only the distributional parameters of Z . When applying a MGF to estimate the ATE, the observed values are used to estimate these distributional parameters of Z . This way, the conditional effects are integrated across the assumed distribution of Z . We call this the moment-based approach.

The moment-based approach is illustrated in Fig. 2.1c. We assume that the z_i are drawn from a normal distribution, that is, $Z \sim \mathcal{N}(\mu, \sigma^2)$. We use the sample mean and variance to estimate $\hat{\mu}$ and $\hat{\sigma}^2$. Using the moment-generating function of the normal distribution,

the moment-based estimator is given by

$$\begin{aligned} \widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) &= \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \exp\left(\hat{\mu}(\hat{\beta}_{01} + \hat{\beta}_{11}) + \frac{\hat{\sigma}^2(\hat{\beta}_{01} + \hat{\beta}_{11})^2}{2}\right) \\ &\quad - \exp(\hat{\beta}_{00}) \exp\left(\hat{\mu}\hat{\beta}_{01} + \frac{\hat{\sigma}^2\hat{\beta}_{01}^2}{2}\right). \end{aligned}$$

In this way, we implicitly integrate over all possible conditional effects of a normally distributed Z and not only the ten observed effects, as is indicated by the shaded area in Fig. 2.1c.

The optimal estimation technique for the distribution parameters $\hat{\theta}$ and their covariance $\mathbf{V}(\hat{\theta})$ depends on the actual distribution of Z and further available information. For example, the method of moments and maximum likelihood estimators often yield identical point estimates, but the MLE is usually more efficient. If prior information about the parameters is available, researchers might consider a Bayesian approach for estimating the parameters. Therefore, an appropriate case-by-case-based decision about an appropriate estimation technique is recommended.

The squared standard error of $\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta})$ is given by

$$\text{SE}_{\text{MOM}}^2 = \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \hat{\beta}}\right)^T \mathbf{V}(\hat{\beta}) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \hat{\beta}} + \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \hat{\theta}}\right)^T \mathbf{V}(\hat{\theta}) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \hat{\theta}}.$$

Here, the covariance of the covariate's moments is required in addition to the covariance of the regression parameters (for proof see "Appendix 2.B"). Otherwise stated, sampling variation for the covariate is taken into account.

2.5 Simulation Study

The goals of our simulation study are twofold. First, we seek to evaluate the statistical performance of our new estimator and check its sensitivity against violations of the distributional assumption. We compare the effect estimates of the empirical approach $\widehat{\text{ATE}}_{\text{EMP}}$, the moment-based approach $\widehat{\text{ATE}}_{\text{MOM}}$, and TEMZ in a quasi-experimental setting. Second, we evaluate the impact of the fixed-covariate assumption compared to the stochastic covariate assumption with regard to standard errors. We examine whether the standard errors based on the fixed-covariate assumption, SE_{EMP} , are underestimated, and whether the stochastic covariate assumption, SE_{MOM} , leads to more accurate standard errors if Z is in fact sampled and not fixed.

An overview of the simulation design is given in Table 2.3. A detailed description of the data generating process can be found in Supplementary Material A. We also provide the corresponding R code in the supplementary material. In all scenarios, data were generated

Table 2.3: Design of the simulation study.

Parameter	Values
Distribution of Z	Uniform $U[-\sqrt{3}, \sqrt{3}]$ Normal $\mathcal{N}(0, 1)$ Poisson $\mathcal{P}(\lambda = 1)$
Effect size Δ of ATE	$-0.5, 0, 0.5$
Standardized interaction effect β_{11}^*	$0, 0.25, 0.5, 0.75, 1$
Sample size N	$50, 125, 250, 500, 1000$

assuming complete data without missing values.²

2.5.1 Data Analysis

Estimation of ATE

We estimated the ATE for every simulated scenario using the empirical estimator $\widehat{\text{ATE}}_{\text{EMP}}$ and the moment-based approach $\widehat{\text{ATE}}_{\text{MOM}}$. Additionally, we estimated the TEMZ with

$$\widehat{\text{TEMZ}}(\hat{\beta}) = \exp(\hat{\beta}_{00} + \hat{\beta}_{10} + \hat{\beta}_{01}\bar{z} + \hat{\beta}_{11}\bar{z}) - \exp(\hat{\beta}_{00} + \hat{\beta}_{01}\bar{z})$$

where $\bar{z} = \frac{1}{N} \sum_{i=1}^N z_i$. According to Eqs. (2.8) and (2.9), the TEMZ should significantly differ from the ATE estimates. For the moment-based approach, we estimated $\widehat{\text{ATE}}_{\text{MOM}}$ with both an accurate and an inaccurate assumption about the distribution of Z . We always chose a normal distribution for the latter. In both procedures, the parameters of the distribution were estimated via the sample moments.³ Sample moments and their covariance were estimated using a multigroup structural equation model with stochastic group sizes. For more information on the procedure, see Supplementary Material A or Mayer et al. (2016)

We examined the bias of the ATE estimate provided by each estimator, i.e., the deviation of the average point estimate from the true ATE. For better comparability across conditions, the bias was standardized using the true standard deviation in the control group. Thus, the bias can be interpreted in effect size units, e.g., a bias of 0.3 indicates that the estimated $\widehat{\text{ATE}}$ overestimates the true effect size by 0.3.

$$B(\widehat{\text{ATE}}) = \frac{\widehat{\text{ATE}} - \text{ATE}}{\sqrt{\text{Var}(Y|X=0)}}.$$

²The data generating process is idealized in the sense that practical problems such as missing data, interference between units, or treatment non-compliance are not taken into account. In the discussion, we provide references on how to deal with such issues in practice.

³Note that MLE would provide equivalent results for a Poisson or normally distributed covariate. For a uniformly distributed covariate, the sample moments-based parameter estimation is unbiased, while the MLE is not necessarily unbiased in small samples.

Note that the true standard deviation $\sqrt{\text{Var}(Y|X=0)}$ was used instead of the estimated sample standard deviation in order to evaluate the ATE bias alone. The true ATE for each condition was computed using Eq. (2.12).

Estimation of Standard Errors

For the empirical approach, we computed SE_{EMP} . For the moment-based approach, the stochastic covariate based SE_{MOM} was computed. Standard errors for $\widehat{\text{TEMZ}}$ were not computed as they are not central to the goal of our simulation study.

First, we examined the coverage of 95% confidence intervals using the critical value $t_{0.975;N-1}$ from a t -distribution. Accuracy of coverage was evaluated using the 90% threshold suggested by Collins, Schafer, and Kam (2001), that is, coverages below this threshold are considered problematic. Finally, we evaluated Type 1 error rates and power.

2.5.2 Results

Estimation of the ATE

As the $\widehat{\text{ATE}}$ point estimates exhibited heavily tailed distributions, especially for small sample sizes, we used the robust location estimator suggested by Huber (1964) to obtain the average point estimate. We chose a Huber M-estimator with an efficiency of 95%, i.e., the average point estimation is close to the sample mean for normally distributed sample statistics, but much more efficient for heavily tailed distributions.⁴ Results are shown in Fig. 2.2.

In the empirical approach, $\widehat{\text{ATE}}_{\text{EMP}}$ bias was between $[-0.66; 0.01]$ and linked to the sample size and interaction effect, especially when the covariate was skewed. Stronger interaction effects led to a larger underestimation of the true average effect. This relation was moderated by sample size, as bias decreased in larger samples. However, in scenarios with a Poisson-distributed covariate and a strong interaction, the bias was surprisingly high, even in large samples (i.e., $N = 500$; $N = 1000$). Although $\widehat{\text{ATE}}_{\text{EMP}}$ is a consistent estimator, it seems to require very large samples to be unbiased under such circumstances.

In the moment-based approach, $\widehat{\text{ATE}}_{\text{MOM}}$ bias was between $[-0.06; 0.33]$ and also linked to the interaction effect, but this was only noticeable in small samples (i.e., $N = 50$; $N = 100$) with a Poisson-distributed covariate. Here, stronger interaction effects led to a larger overestimation of the true average effect. This bias decreased quickly with increasing sample size; thus, $\widehat{\text{ATE}}_{\text{MOM}}$ was unbiased for medium to large samples.

Unsurprisingly, this unbiasedness of $\widehat{\text{ATE}}_{\text{MOM}}$ depended on an accurate selection of the moment-generating function. In settings where the normality assumption for Z was violated, the bias increased with the interaction effect and was between $[-0.92; 0.39]$. Nonetheless, in scenarios with no interaction effect, the violation did not lead to bias.

⁴The Huber M-estimator is similar to a winsorized mean, that is, parts of the sample at the low and high ends are replaced with the most extreme remaining values. The low and high ends are determined using the median and a corrected median absolute deviation. For details, see Huber and Ronchetti (2009).

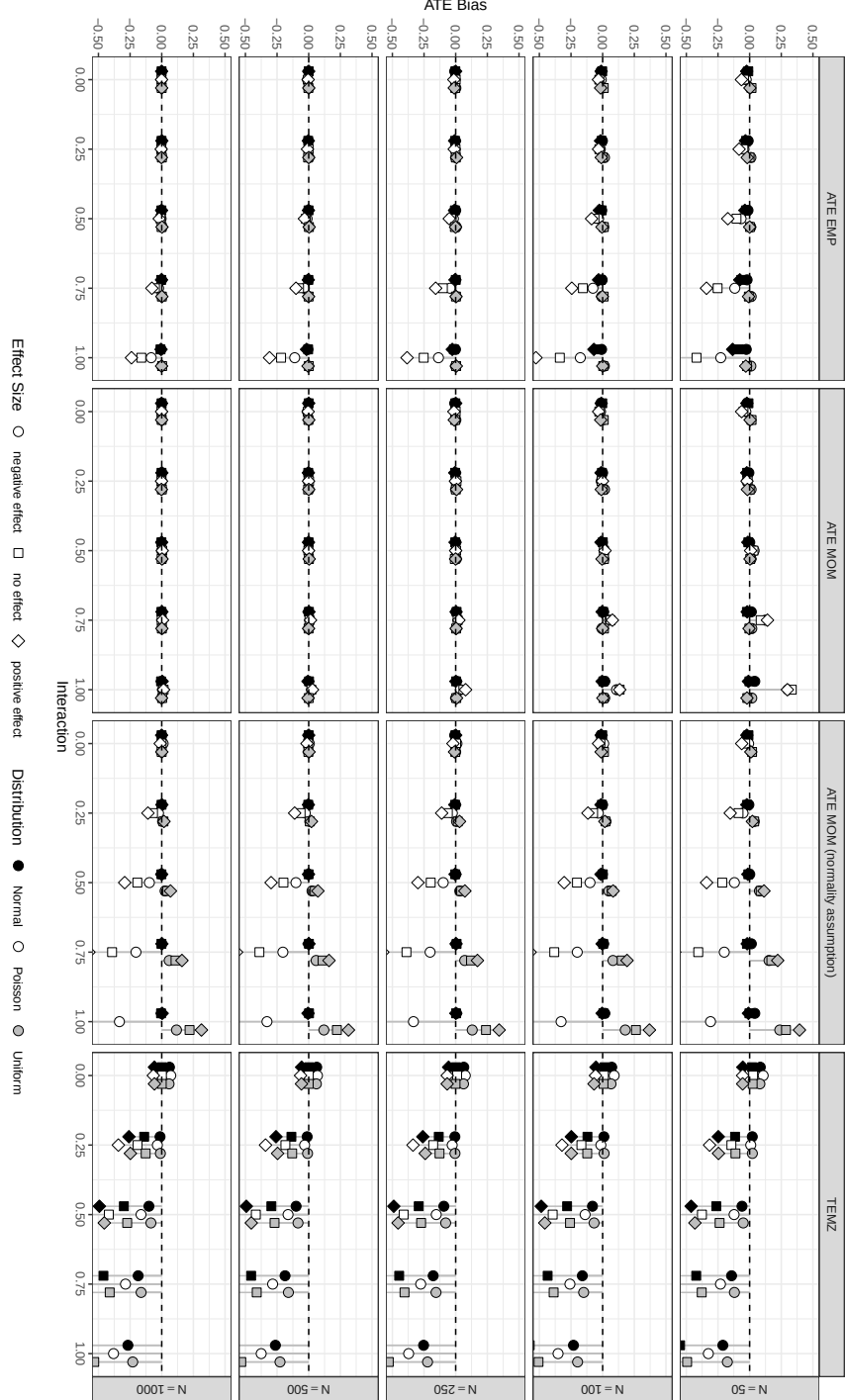


Figure 2.2

Bias $B(\widehat{ATE})$ for the estimators $\widehat{ATE}_{EMP}$ (ATE EMP), $\widehat{ATE}_{MOM}$ (ATE MOM), and $\widehat{TE}_{MZ}$ (TENZ). For $\widehat{ATE}_{MOM}$, an additional column is shown where the unconditional distribution of Z was always assumed to be normally distributed. Dashed line at $B(\widehat{ATE}) = 0$ indicates no bias. Magnitude of bias can be interpreted as a standardized effect size like Cohen's d .

Our results suggest that the TEMZ is only a valid estimate of the ATE when there is both no interaction and no average treatment effect. In most other scenarios, the TEMZ is substantively smaller than the ATE and the bias in the ATE estimate was between $[-1.24; 0.11]$.

Estimation of Standard Errors

First, we evaluated the coverage of 95% confidence intervals based on the different standard error estimates. Results are shown in Fig. 2.3. For $\widehat{SE}_{EMP}$, coverage yielded mixed results between $[0.53; 0.96]$. Again, with a skewed covariate and strong interaction confidence interval's coverage decreased to 53%. The same decrease occurred for a uniformly distributed covariate, but to a lesser extent. Increases in sample size seemed to foster inaccurate coverage. The coverage of CIs based on $\widehat{SE}_{MOM}$ ($[0.91; 0.99]$) was shown to be accurate under all scenarios, that is, coverage was always above the 90% threshold.

Next, we examined Type 1 error rates and power when making inferences based on $\widehat{SE}_{MOM}$ and $\widehat{SE}_{EMP}$. The results are shown in Fig. 2.4. As expected, the Type 1 error rates for $\widehat{SE}_{EMP}$ were between $[0.04; 0.40]$. For a skewed covariate and strong interaction, the Type 1 error rate increased to 40%. The moment-based approach yielded accurate Type 1 error rates (between $[0.01; 0.06]$). However, in small samples the Type 1 error rates for the moment-based approach tended to be too low.

2.6 Empirical Illustration

We use data from an intervention study for problem drinking (Muench et al., 2017) to illustrate the approach presented in the previous section. In a randomized controlled trial, Muench et al. (2017) examined whether mobile messaging interventions reduce problem drinking. We focus on comparing the most effective intervention (tailored adaptive intervention, $X = 1$, $N = 31$) to the control group (mobile assessment only, $X = 0$, $N = 28$). The primary outcome Y is a count variable, indicating the weekly sum of standard drinks after 12 weeks ranging from 0 to 43. The baseline measure for weekly sum of standard drinks Z_{pre} ranging from 12 to 50 is included as covariate in our analysis (control group: $M = 24.69$, $SD = 9.81$; treatment group: $M = 24.16$, $SD = 6.46$).

In the original study, a Poisson GLM was used with a scaling parameter ϕ allowing for overdispersion. Muench et al. (2017) did not include interaction terms in the model. In our analysis, we also used a Poisson GLM with scaling parameter and added a parameter β_{11} for the interaction between the treatment variable and the covariate. The covariate Z_{pre} was grand mean centered with respect to the mean of the full sample including all five treatment groups. For our subsample, the mean of the covariate was $\bar{z} = -0.467$. For the moment-based approach, we used the non-centered version of the covariate, because the Poisson distribution and the negative binomial distribution cannot meaningfully be

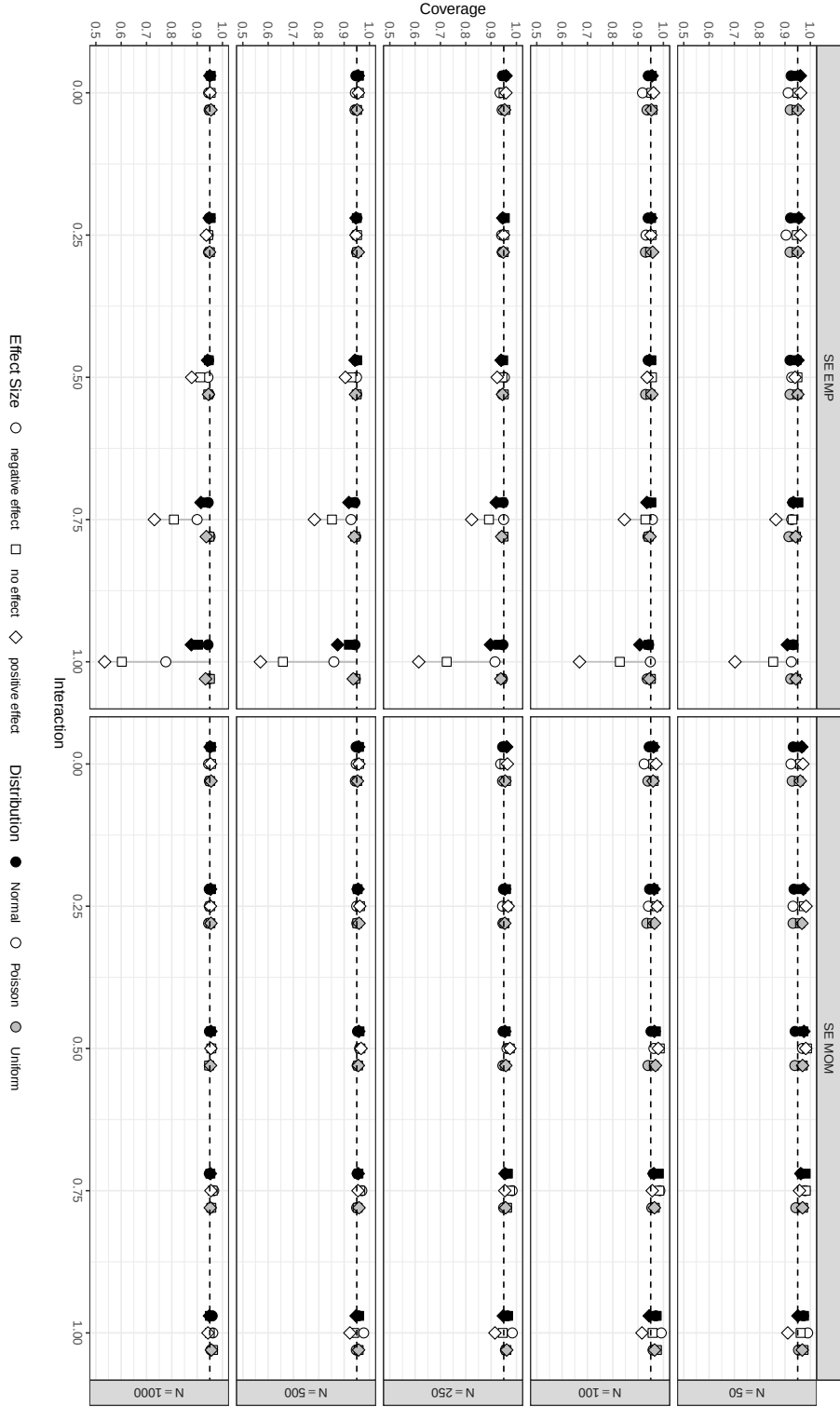


Figure 2.3

Coverage of 95% confidence intervals based on the respective standard error estimates SE_{EMP} (SE EMP) and SE_{MOM} (SE MONI). That is, the percentage of confidence intervals that include the true value of the average treatment effect. Dashed line indicates the true value.

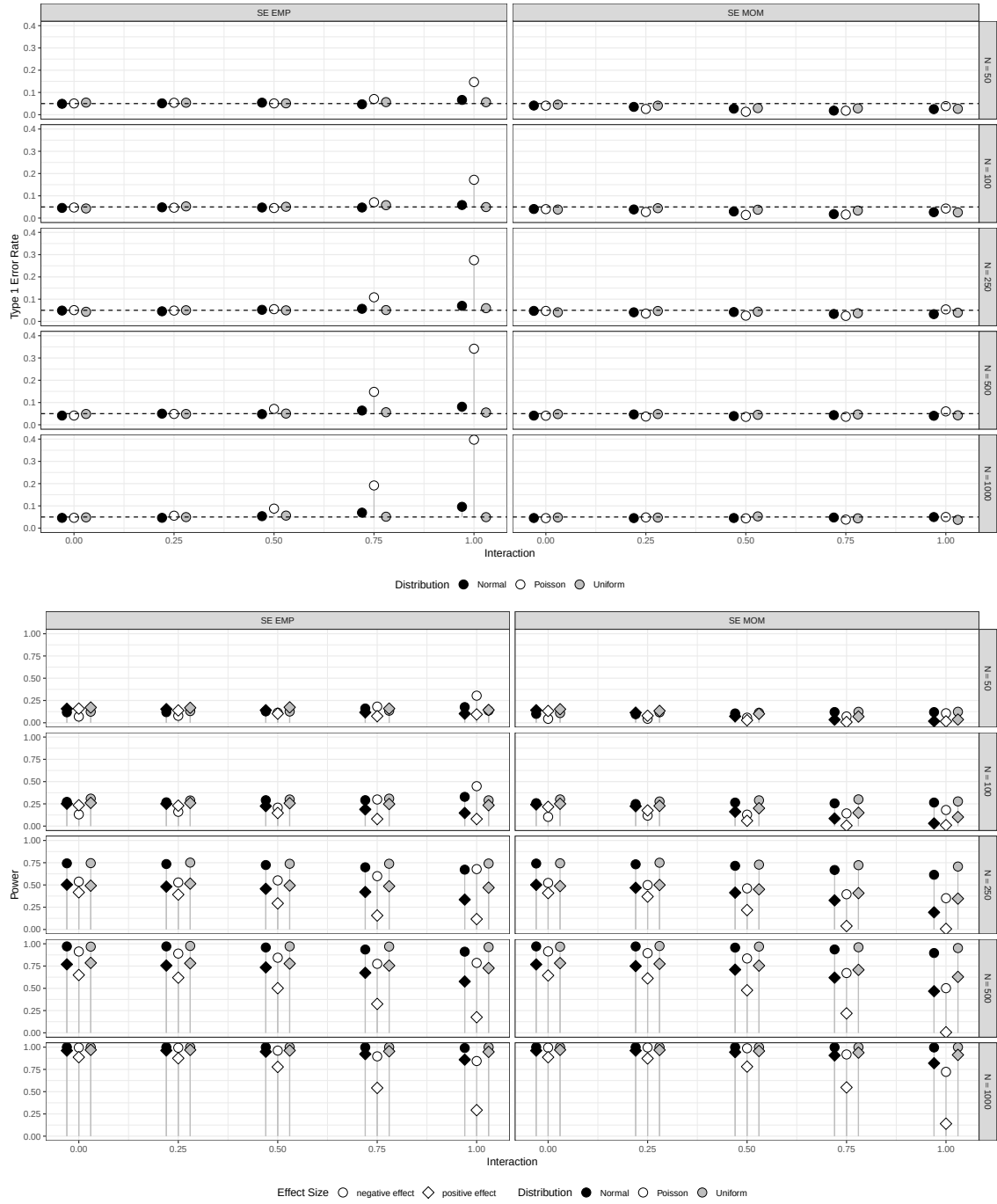


Figure 2.4

Plots of Type 1 error rates and power for standard error $\widehat{ATE}_{EMP}$ (SE EMP) and standard error for $\widehat{ATE}_{MOM}$ (SE MOM). **a** Type 1 error rates are shown, that is, the percentage of 95% CIs that do not include the value 0 when the true average effect is zero. As indicated by the dashed line, these error rates are expected to be about 5%. **b** The power is shown, that is, the percentage of 95% CIs that do not include the value 0 when the true average effect is not zero.

centered around the zero point.⁵

Moment-Based Estimator In order to formulate the moment-based estimator, the distribution of the covariate has to be determined. The covariate sum of standard drinks at baseline Z_{pre} is a count variable like the outcome variable. Tests for distribution of Z_{pre} revealed that a Poisson distribution ($\chi^2(50) = 10424$, $p < 0.001$) cannot be assumed, but a negative binomial distribution ($\chi^2(50) = 42.974$, $p = 0.749$) was reasonable. Hence, we modeled $Z_{\text{pre}} \sim \text{NB}(r, p)$ and the moment-based estimator is given by

$$\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) = \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \left(\frac{\hat{p} \exp(\hat{\beta}_{01} + \hat{\beta}_{11})}{1 - (1 - \hat{p}) \exp(\hat{\beta}_{01} + \hat{\beta}_{11})} \right)^{\hat{r}} - \exp(\hat{\beta}_{00}) \left(\frac{\hat{p} \exp(\hat{\beta}_{01})}{1 - (1 - \hat{p}) \exp(\hat{\beta}_{01})} \right)^{\hat{r}}$$

where $\hat{r} = \bar{z}^2 / (\bar{z} + s_z^2)$ and $\hat{p} = \bar{z} / (\bar{z} + s_z^2)$ with sample mean $\bar{z}$ and sample variance s_z^2 .

Results The Poisson regression model showed an adequate fit with a pseudo- $R^2 = 39.2\%$ and scaling parameter $\phi = 2.384$. The estimate for the intercept was $\hat{\beta}_{00} = 3.090$, $p < 0.001$, $\text{CI} = [2.965; 3.216]$, which means that for a person in the control group with an average alcohol consumption at baseline ($Z_{\text{pre}} = 0$), the expected outcome is $\hat{E}(Y|X = 0, Z = 0) = \exp(3.090) = 22.042$ weekly standard drinks. The estimated regression coefficient of Z_{pre} in the control group was $\hat{\beta}_{01} = 0.029$, $p < 0.001$, $\text{CI} = [0.016; 0.042]$, that is, for a 1-unit increase in Z_{pre} , the expected count is multiplied by $\exp(0.029) = 1.029$. So in the control group, each additional drink at baseline is associated with a 3 percent increase in the expected weekly drinks at the posttest. The interaction effect was not significant ($\hat{\beta}_{11} = -0.010$, $p = 0.471$, $\text{CI} = [-0.036; 0.017]$), indicating that the effect of the treatment does not depend on the values of the pretest and vice versa.

In order to assess the effectiveness of the tailored adaptive intervention, we first consider the regression coefficient for the treatment⁶ $\hat{\beta}_{10} = -0.402$, $p < 0.001$, $\text{CI} = [-0.594; -0.209]$. Exponentiating the regression coefficient yields $\exp(-0.402) = 0.67$, that is, the number of weekly drinks was significantly reduced about one third compared to the control group. When we exponentiate $\hat{\beta}_{10}$, we obtain a treatment effect defined as ratio of conditional expectations. The treatment effect then represents the ratio of expected counts in the treatment group compared to the control group at the grand mean $Z_{\text{pre}} = 0$. We can also consider effects defined as differences between conditional expectations in this example. For this purpose, we first compute the expected outcome in the treatment group $\hat{E}(Y|X = 1, Z = 0) = 14.729$ and subtract the expected outcome in the control group $\hat{E}(Y|X = 0, Z = 0) = 22.042$ which gives the TEMZ = -7.313 as defined in Eq.

⁵As a supplement to this article, we provide an R package ([R Core Team, 2018](#)) called CountEffects. It is available from GitHub and can be installed as usual (see Supplementary Material A for details).

⁶[Muench et al. \(2017\)](#) reported $\hat{\beta}_1 = -0.40$ based on their model without interaction.

(2.7). The TEMZ represents the difference between the expected counts at the grand mean of Z_{pre} . So, an average person is expected to consume 7.313 weekly drinks less in the treatment group compared to the control group. In a similar vein, we can compute the treatment effect at the subsample-specific mean $\text{TEMZ}^{(s)} = -7.106$, $\text{SE} = 1.734$, $\Delta = -0.736$, which corresponds to the difference between expected counts at $\bar{z} = -0.467$ and is slightly smaller than the effect at the grand mean.

Next, we consider the average treatment effect as defined in Eq. (2.6) and compare the empirical approach and the newly proposed moment-based approach with stochastic covariates. The empirical approach yields $\widehat{\text{ATE}}_{\text{EMP}} = -7.263$, $\text{SE}_{\text{EMP}} = 1.725$, $\Delta = -0.752$. It gives an estimate for the difference between expected counts in the treatment group and the control group averaged over all observed values of Z_{pre} in the sample, which corresponds to the difference between $\hat{E}[E(Y|X=0, Z)] = 21.824$ and $\hat{E}[E(Y|X=1, Z)] = 14.561$. In the moment-based approach, these expected counts are estimated based on the moment-generating function and hence differ slightly from the empirical equivalents: $\hat{E}[E(Y|X=0, Z)] = 22.328$ and $\hat{E}[E(Y|X=1, Z)] = 14.768$. The difference between these expected counts is the ATE using the moment-based estimator: $\widehat{\text{ATE}}_{\text{MOM}} = -7.560$, $\text{SE}_{\text{MOM}} = 1.810$, $\Delta = -0.783$. So, on average, persons are expected to consume 7.560 weekly drinks less in the treatment group compared to the control group.

In this example, the empirical and the moment-based approaches provide similar results both with respect to point estimates and standard errors. This is not surprising and in line with our simulations, because the (standardized) interaction effect $\hat{\beta}_{11}$ was close to zero in this empirical illustration. In our simulations, we have demonstrated that in scenarios with no or low interaction both approaches are reliable.

2.7 Discussion

In this paper, we developed and evaluated a new method of estimating average treatment effects for count data using stochastic covariates. The new moment-based method uses moment-generating functions to account for the distribution of the covariate and allows the ATE to be computed analytically rather than approximated numerically. We found that the fixed-covariate assumption in the empirical approach leads to an underestimation of standard errors for the average treatment effect. The moment-based approach provided accurate effect and standard error estimates and outperformed the empirical approach when a skewed covariate and/or strong interaction was present. However, an appropriate choice of moment-generating function is crucial for the accurateness of the moment-based approach.

The standard errors in the empirical approach were substantively too small in some scenarios due to the fixed-covariate assumption. The coverage of the corresponding confidence intervals decreased from the required 95% to a very low value of 53%. As a result, inferences made on the basis of the ATE estimates can be problematic in the empirical approach. The reliability of inferences could be improved by using an estimator that allows

for stochastic covariates (i.e., the moment-based approach). These results are in line with previous findings on the fixed-covariate assumption in linear models (Chen, 2006; Kröhne, 2009; Y. Liu et al., 2017).

The moment-based and empirical approaches are consistent with regard to the point estimation of the ATE. However, the moment-based approach requires substantially smaller samples to deliver unbiased results than the empirical approach. The TEMZ, an appropriate substitute for the ATE in linear models, is strongly biased in regressions with log link. Our simulation results indicate that the TEMZ often substantively underestimates the ATE and should therefore not be used as a measure of average effectiveness.

The derivations in our paper, especially the presentation of the moment-based approach, are limited to a single covariate. There are different possibilities to extend our new approach to multiple covariates. First, the linear component in a regression with log link formally represents a weighted sum of random variables that results in an unidimensional random variable. Hence, the moment-based estimator can be formulated assuming a specific distribution for the resulting random variable. Second, moment-generating functions exist for few multidimensional distributions such as the multivariate normal distribution. Thus, it is possible to specify a multivariate moment-based estimator under the assumption of multivariate normality.

We introduced average effect estimation as a two-step procedure and applied basic approaches to estimating these two steps (i.e., maximum likelihood for NB-GLM, sample moments for moment-based approach). We pointed out that each step can be estimated via several methods. In addition, the two steps can be estimated with different estimation techniques. For example, if there is prior information about the regression coefficients, researchers might apply a Bayesian approach to estimate the regression model, but a conventional maximum likelihood approach for the parameters of the distribution. That is, the implementation of the presented approach is very flexible. For some estimation approaches (e.g., Bayesian, maximum likelihood), it is possible to estimate the two-step procedure simultaneously. However, as we have shown, the estimation of the two steps is independent. Thus, we would not expect any benefit from a simultaneous estimation procedure.

Our paper is limited to the derivation of average treatment effects based on differences. We would expect our findings to apply to other kinds of effects that also rely on a fixed-covariate assumption, such as effects based on ratios or conditional treatment effects for categorical covariates. Conditional treatment effects for categorical covariates are examined when the effects for specific subgroups are of interest (e.g., average effect on the treated or average effect for women). If the covariates are in fact sampled from a population and this sampling variability is neglected in the estimation process, the resulting standard errors might be deflated. In contrast, for procedures that seek to estimate the conditional effect for fixed values of a continuous covariate (e.g., J-N procedure by Johnson & Neyman, 1936), when an a priori fixed value (e.g., 140 mm Hg to denote hypertension) is utilized, the fixed covariate assumption is adequate. However, when sample-based val-

ues are utilized the fixed-covariate assumption might lead to problematic J–N significance tests and confidence intervals.

A causal interpretation of the ATE is possible if additional assumptions hold. For our example, a temporal order has to be implied, that is, the treatment X precedes the outcome Y , and the covariate Z is prior or simultaneous to the treatment X . Throughout the paper, for instance, we mentioned the independence of X and Z as a causality condition for the ATE which can be obtained with randomized assignment to the treatment groups. Other causality conditions can be found in [Mayer, Thoemmes, et al. \(2014\)](#). In applied settings, studies oftentimes suffer from practical problems. In this case, an analytic strategy that can adequately deal with problems such as missing data (see, e.g., [Graham, 2012](#); [Little & Rubin, 2002](#); [Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke, 2014](#)), interference between units (e.g., [Imbens & Rubin, 2015](#); [Rosenbaum, 2007](#)), or treatment non-compliance (e.g., [Sagarin et al., 2014](#); [Shadish et al., 2002](#)) has to be chosen.

In conclusion, the examination of (average) treatment effects is an important task for (quasi-)experimental psychologists. While point estimation of these effects often seems straightforward, the prerequisites for inference have to be carefully considered. In recent years, the development of statistical software (e.g., EffectLiteR, [Mayer et al. 2016](#)) has increasingly allowed researchers to make explicit decisions about these prerequisites in linear models, and we hope to contribute by creating such possibilities for nonlinear models.

Appendix

2.A Non-Collapsibility of Ratio Effects

Theorem 1. *Let Y be a nonnegative, X a binary, and Z a continuous random variable with $\text{Var}(Z) > 0$ and*

$$E(Y|X, Z) = \exp(\beta_{00} + \beta_{10}X + \beta_{01}Z + \beta_{11}XZ) \quad (2.13)$$

where $\beta_{00}, \beta_{10}, \beta_{01}, \beta_{11} \in \mathbb{R}$.

The treatment effect based on a ratio is collapsible, that is,

$$E \left[\frac{E(Y|X=1, Z)}{E(Y|X=0, Z)} \right] = \frac{E(E(Y|X=1, Z))}{E(E(Y|X=0, Z))}$$

the expectation across $Z = z$ -conditional ratios is equal to the ratio of Z -adjusted group expectations if and only if

$$\beta_{01} = 0 \vee \beta_{11} = 0,$$

that is, either no covariate effect in control group (i.e., $\beta_{01} = 0$) or no interaction effect (i.e., $\beta_{11} = 0$). Thus, in all other cases, the ratio-based treatment effect is not collapsible.

Proof. We start at

$$E \left(\frac{E(Y|X=1, Z)}{E(Y|X=0, Z)} \right) = \frac{E(E(Y|X=1, Z))}{E(E(Y|X=0, Z))} \quad (2.14)$$

Then, the parameterized form of the conditional expectation $E(Y|X, Z)$ given in Eq. (2.13) is inserted into Eq. (2.14):

$$\begin{aligned} E \left(\frac{\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z)}{\exp(\beta_{00} + \beta_{01}Z)} \right) &= \frac{E(\exp(\beta_{00} + \beta_{10} + \beta_{01}Z + \beta_{11}Z))}{E(\exp(\beta_{00} + \beta_{01}Z))} \\ \Leftrightarrow \frac{\exp(\beta_{00} + \beta_{10})}{\exp(\beta_{00})} E \left(\frac{\exp(\beta_{01}Z + \beta_{11}Z)}{\exp(\beta_{01}Z)} \right) &= \frac{\exp(\beta_{00} + \beta_{10})}{\exp(\beta_{00})} \frac{E(\exp(\beta_{01}Z + \beta_{11}Z))}{E(\exp(\beta_{01}Z))} \end{aligned} \quad (2.15)$$

To obtain Eq. (2.15) we used the linearity property of expectations. Now, we apply basic transformations and obtain

$$\begin{aligned} E(\exp(\beta_{11}Z)) &= \frac{E(\exp(\beta_{01}Z + \beta_{11}Z))}{E(\exp(\beta_{01}Z))} \\ \Leftrightarrow E(\exp(\beta_{11}Z)) E(\exp(\beta_{01}Z)) &= E(\exp(\beta_{01}Z) \exp(\beta_{11}Z)). \end{aligned}$$

In general,

$$\text{Cov}(\exp(\beta_{01}Z), \exp(\beta_{11}Z)) = E(\exp(\beta_{01}Z) \exp(\beta_{11}Z)) - E(\exp(\beta_{11}Z)) E(\exp(\beta_{01}Z))$$

holds, and thus Equation (2.A) implies that

$$\text{Cov}(\exp(\beta_{01}Z), \exp(\beta_{11}Z)) = 0.$$

There are two solutions to this equation, namely

$$\beta_{01} = 0 \vee \beta_{11} = 0.$$

□

2.B Standard Error of Moment-Based Estimator

In this section, we show how the standard error for the moment-based approach with stochastic covariates is obtained. The presentation is divided in two parts. First, the general standard error formula for the moment-based approach is derived using the delta method. Second, we show that the limiting covariance $\mathbf{V}(\beta_0, \theta_0)$ is characterized by an independence of β_0 and θ_0 .

Let Y be a nonnegative random variable, X be a binary random variable, and Z a continuous real-valued random variable. Consider a sample of independent observations (Y_i, X_i, Z_i) , $i = 1, \dots, N$, where N is the sample size. Then, the estimator $\widehat{\text{ATE}}_{\text{MOM}}$ is given by

$$\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) = \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \widehat{M}_Z(\hat{\beta}_{01} + \hat{\beta}_{11}) - \exp(\hat{\beta}_{00}) \widehat{M}_Z(\hat{\beta}_{01}) \quad (2.16)$$

where $\hat{\beta} = (\hat{\beta}_{00}, \hat{\beta}_{10}, \hat{\beta}_{01}, \hat{\beta}_{11})$ and $\hat{\theta}$ a vector of sample moments. These can be central or raw moments.

Theorem 2. *The function $\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta})$ is given in Eq. (2.16). Then, its asymptotic variance (i.e., squared standard error) is given by*

$$\text{SE}_{\text{MOM}}^2 = \left(\nabla \widehat{\text{ATE}}_{\text{MOM}} \right)^T \mathbf{V}(\beta_0, \theta_0) \nabla \widehat{\text{ATE}}_{\text{MOM}}, \quad (2.17)$$

where $\mathbf{V}(\beta_0, \theta_0)$ is the asymptotic covariance of the parameters β_0 and θ_0 .

Proof. As $\widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta})$ is a real-valued function of a vector, the delta theorem can be applied. This immediately delivers Eq. (2.17). For more information and proof, see Theorem 5.19 in Boos and Stefanski (2013) or Raykov and Marcoulides (2004, p. 628). □

In the next step, we examine the structure of the asymptotic covariance $\mathbf{V}(\beta_0, \theta_0)$. First, we show how the estimators for $\hat{\beta}$ and $\hat{\theta}$ can be notated as M-estimators. We then use the M-estimation framework to derive the asymptotic covariance $\mathbf{V}(\beta_0, \theta_0)$. For more details on M-estimation, see Boos and Stefanski (2013, Chapter 7)

The parameter estimates $\hat{\beta}$ and $\hat{\theta}$ are solutions for a vector equation

$$\sum_{i=1}^N \psi(\epsilon_i, X_i, Z_i, \hat{\beta}, \hat{\theta}) = \mathbf{0},$$

with functions ψ :

$$\psi_1(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = \epsilon_1, \quad \text{for } \beta_{00} \quad (2.18)$$

$$\psi_2(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = \epsilon_1 \cdot X_1, \quad \text{for } \beta_{10}$$

$$\psi_3(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = \epsilon_1 \cdot Z_1, \quad \text{for } \beta_{01}$$

$$\psi_4(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = \epsilon_1 \cdot X_1 Z_1, \quad \text{for } \beta_{11}$$

$$\psi_{4+1}(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = Z_1^{m_1} - \theta_1, \quad \text{for } \theta_1$$

$$\vdots = \vdots$$

$$\psi_{4+n}(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) = Z_1^{m_p} - \theta_p, \quad \text{for } \theta_p \quad (2.19)$$

where

$$\epsilon_i = Y_i - \exp(\beta_{00} + \beta_{10}X_i + \beta_{01}Z_i + \beta_{11}X_iZ_i)$$

and thus are called M-estimators. That is, we can derive the asymptotic covariance $\mathbf{V}(\beta_0, \theta_0)$ via the M-estimation framework.

Theorem 3. *The asymptotic covariance matrix $\mathbf{V}(\beta_0, \theta_0)$ is*

$$\mathbf{V}(\beta_0, \theta_0) = \begin{pmatrix} \mathbf{V}(\beta_0) & 0 \\ 0 & \mathbf{V}(\theta_0) \end{pmatrix}$$

where $\mathbf{V}(\beta_0)$ is the asymptotic covariance matrix of β_0 and $\mathbf{V}(\theta_0)$ is the asymptotic covariance matrix of θ_0 . That is, the estimates $\hat{\beta}$ and $\hat{\theta}$ are independent.

Proof. The asymptotic covariance $\mathbf{V}(\beta_0, \theta_0)$ can be decomposed to

$$\mathbf{V}(\beta_0, \theta_0) = \mathbf{A}(\beta_0, \theta_0)^{-1} \mathbf{B}(\beta_0, \theta_0) (\mathbf{A}(\beta_0, \theta_0)^{-1})^T$$

The asymptotic covariance $\mathbf{V}(\beta_0, \theta_0)$ is often called sandwich matrix because the “meat” $\mathbf{B}(\beta_0, \theta_0)$ is placed between the “bread” $\mathbf{A}(\beta_0, \theta_0)^{-1}$ and $(\mathbf{A}(\beta_0, \theta_0)^{-1})^T$ (cf. [Boos & Stefanski, 2013](#), p. 301).

The “bread” matrix $\mathbf{A}(\beta_0, \theta_0)$ The matrix $\mathbf{A}(\beta_0, \theta_0)$ is defined as

$$\mathbf{A}(\beta_0, \theta_0) = E\{-\psi'(\epsilon_1, X_1, Z_1, \beta_0, \theta_0)\},$$

that is, the partial derivatives of the ψ_i -functions with respect to β_0 and θ_0 have to be considered.

As the parameters θ_0 are not included in the functions ψ_1 to ψ_4 and vice versa, the parameters β_0 are not included in the functions ψ_{4+1} to ψ_{4+p} ; it is trivial to see that the corresponding partial derivatives are equal to zero. Thus, the resulting matrix is block diagonal:

$$\mathbf{A}(\beta_0, \theta_0) = \begin{pmatrix} \mathbf{A}(\beta_0) & 0 \\ 0 & \mathbf{A}(\theta_0) \end{pmatrix}$$

Using properties of block diagonal matrices, $\mathbf{A}(\beta, \theta)^{-1}$ can be computed as

$$\mathbf{A}(\beta, \theta)^{-1} = \begin{pmatrix} \mathbf{A}(\beta)^{-1} & 0 \\ 0 & \mathbf{A}(\theta)^{-1} \end{pmatrix} \quad (2.20)$$

The “meat” matrix $\mathbf{B}(\beta_0, \theta_0)$ The matrix $\mathbf{B}(\beta_0, \theta_0)$ is defined as

$$\mathbf{B}(\beta_0, \theta_0) = E\{\psi(\epsilon_1, X_1, Z_1, \beta_0, \theta_0) \psi(\epsilon_1, X_1, Z_1, \beta_0, \theta_0)^T\}.$$

For the Eqs. (2.18) to (2.19) this yields

$$\mathbf{B}(\beta_0, \theta_0) = E \begin{pmatrix} \psi_1^2 & \psi_1\psi_2 & \psi_1\psi_3 & \psi_1\psi_4 & \psi_1\psi_{4+1} & \dots & \psi_1\psi_{4+p} \\ \psi_2\psi_1 & \psi_2^2 & \psi_2\psi_3 & \psi_2\psi_4 & \psi_2\psi_{4+1} & \dots & \psi_2\psi_{4+p} \\ \psi_3\psi_1 & \psi_3\psi_2 & \psi_3^2 & \psi_3\psi_4 & \psi_3\psi_{4+1} & \dots & \psi_3\psi_{4+p} \\ \psi_4\psi_1 & \psi_4\psi_2 & \psi_4\psi_3 & \psi_4^2 & \psi_4\psi_{4+1} & \dots & \psi_4\psi_{4+p} \\ \psi_{4+1}\psi_1 & \psi_{4+1}\psi_2 & \psi_{4+1}\psi_3 & \psi_{4+1}\psi_4 & \psi_{4+1}^2 & \dots & \psi_{4+1}\psi_{4+p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \psi_{4+p}\psi_1 & \psi_{4+p}\psi_2 & \psi_{4+p}\psi_3 & \psi_{4+p}\psi_4 & \psi_{4+p}\psi_{4+1} & \dots & \psi_{4+p}^2 \end{pmatrix}.$$

It can be shown that

$$E[\psi_{4+i}\psi_l] = E[\psi_l\psi_{4+i}] = 0, \quad \forall i = 1, \dots, p, \quad l = 1, 2, 3, 4.$$

For the estimating equations of type $E[\psi_1\psi_{4+i}]$, we have

$$\begin{aligned} E[\psi_1\psi_{4+i}] &= E[(Z_1^{m_i} - \theta_i) \cdot \epsilon_1] \\ &= E[Z_1^{m_i} \cdot \epsilon_1 - \theta_i \cdot \epsilon_1] \\ &= E[Z_1^{m_i} \cdot \epsilon_1] - E[\theta_i \cdot \epsilon_1] \\ &= E[Z_1^{m_i} \cdot \epsilon_1] - \theta_i \cdot E[\epsilon_1] \\ &= E[\epsilon_1] \cdot E[Z_1^{m_i}] - \theta_i E[\epsilon_1], \quad \text{as } \text{Cov}(\epsilon, Z_1^{m_i}) = 0 \\ &= 0 \cdot E[Z_1^{m_i}] - 0 \cdot \theta_i, \quad \text{as } E(\epsilon) = 0 \\ &= 0, \end{aligned}$$

and for the estimating equations of type $E[\psi_2\psi_{4+i}]$, we have

$$\begin{aligned}
 E[\psi_{4+i}\psi_2] &= E[(Z_1^{m_i} - \theta_i) \cdot \epsilon_1 X_1] \\
 &= E[Z_1^{m_i} \cdot \epsilon_1 X_1 - \theta_i \cdot \epsilon_1 X_1] \\
 &= E[Z_1^{m_i} \cdot \epsilon_1 X_1] - E[\theta_i \cdot \epsilon_1 X_1] \\
 &= E[Z_1^{m_i} \cdot \epsilon_1 X_1] - \theta_i \cdot E[\epsilon_1 X_1] \\
 &= E[\epsilon_1] \cdot E[Z_1^{m_i} X_1] - \theta_i E[\epsilon_1] E[X_1], \quad \text{as } \text{Cov}(\epsilon_1, Z_1^{m_i} X_1) = \text{Cov}(\epsilon_1, X_1) = 0 \\
 &= 0 \cdot E[Z_1^{m_i} X_1] - 0 \cdot \theta_i E[X_1], \quad \text{as } E(\epsilon) = 0 \\
 &= 0.
 \end{aligned}$$

The proof is analogous for $E[\psi_3\psi_{4+i}]$ and $E[\psi_4\psi_{4+i}]$ by substituting X_1 with Z_1 or $X_1 Z_1$, respectively.

Thus, the matrix $\mathbf{B}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ is block diagonal:

$$\mathbf{B}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \begin{pmatrix} \mathbf{B}(\boldsymbol{\beta}_0) & 0 \\ 0 & \mathbf{B}(\boldsymbol{\theta}_0) \end{pmatrix}. \quad (2.21)$$

The asymptotic covariance $\mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ As stated above, the asymptotic covariance $\mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ can be obtained via

$$\mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \mathbf{A}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1} \mathbf{B}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) (\mathbf{A}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)^{-1})^T$$

We use our results from Eqs. (2.20) and (2.21) to obtain

$$\mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \begin{pmatrix} \mathbf{A}(\boldsymbol{\beta}_0)^{-1} & 0 \\ 0 & \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{B}(\boldsymbol{\beta}_0) & 0 \\ 0 & \mathbf{B}(\boldsymbol{\theta}_0) \end{pmatrix} \begin{pmatrix} \mathbf{A}(\boldsymbol{\beta}_0)^{-1} & 0 \\ 0 & \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \end{pmatrix}^T$$

At last, basic matrix algebra yields

$$\begin{aligned}
 \mathbf{V}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) &= \begin{pmatrix} \mathbf{A}(\boldsymbol{\beta}_0)^{-1} \mathbf{B}(\boldsymbol{\beta}_0) (\mathbf{A}(\boldsymbol{\beta}_0)^{-1})^T & 0 \\ 0 & \mathbf{A}(\boldsymbol{\theta}_0)^{-1} \mathbf{B}(\boldsymbol{\theta}_0) (\mathbf{A}(\boldsymbol{\theta}_0)^{-1})^T \end{pmatrix} \\
 &= \begin{pmatrix} \mathbf{V}(\boldsymbol{\beta}_0) & 0 \\ 0 & \mathbf{V}(\boldsymbol{\theta}_0) \end{pmatrix}.
 \end{aligned}$$

□

The results of Theorems 2 and 3 directly implicate the following lemma.

Lemma 1. *The asymptotic variance for the function $\widehat{\text{ATE}}_{\text{MOM}}$ is given by*

$$\text{SE}_{\text{MOM}}^2 = \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\beta}_0} \right)^T \mathbf{V}(\boldsymbol{\beta}_0) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\beta}_0} + \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\theta}_0} \right)^T \mathbf{V}(\boldsymbol{\theta}_0) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\theta}_0}.$$

Proof. Combining Theorems 2 and 3, the asymptotic variance of $\widehat{\text{ATE}}_{\text{MOM}}$ is

$$\text{SE}_{\text{MOM}}^2 = \left(\nabla \widehat{\text{ATE}}_{\text{MOM}} \right)^T \begin{pmatrix} \mathbf{V}(\boldsymbol{\beta}_0) & 0 \\ 0 & \mathbf{V}(\boldsymbol{\theta}_0) \end{pmatrix} \nabla \widehat{\text{ATE}}_{\text{MOM}}.$$

The gradient $\nabla \widehat{\text{ATE}}_{\text{MOM}}$ can be rewritten as

$$\begin{aligned} \nabla \widehat{\text{ATE}}_{\text{MOM}} &= \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)} \\ &= \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\beta}_0)}, \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\theta}_0)} \right)^T \end{aligned}$$

which then yields

$$\begin{aligned} \text{SE}_{\text{MOM}}^2 &= \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\beta}_0)}, \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\theta}_0)} \right) \begin{pmatrix} \mathbf{V}(\boldsymbol{\beta}_0) & 0 \\ 0 & \mathbf{V}(\boldsymbol{\theta}_0) \end{pmatrix} \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\beta}_0)}, \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial (\boldsymbol{\theta}_0)} \right)^T \\ &= \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\beta}_0} \right)^T \mathbf{V}(\boldsymbol{\beta}_0) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\beta}_0} + \left(\frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\theta}_0} \right)^T \mathbf{V}(\boldsymbol{\theta}_0) \frac{\partial \widehat{\text{ATE}}_{\text{MOM}}}{\partial \boldsymbol{\theta}_0}. \end{aligned}$$

□

2.C Supplementary Material: Simulation Details

In this section we provide additional information on the simulation study. First, a detailed description of the data generation process is given. Second, we state the explicit moment-based estimators for each unconditional distribution of Z used in the simulation. Third, we provide the listings for the data generation procedure as well as for the estimators.

Data Simulation

As the population model, we choose a regression with log link according to Equation (9) and conditional NB distributions for Y . The model includes a dichotomous treatment variable X , a covariate Z and their interaction XZ .

In order to generate scenarios that are comparable across different distributions of Z , we computed the regression coefficients $\beta_{00}, \beta_{01}, \beta_{11}$ from their standardized versions $\beta_{00}^*, \beta_{01}^*, \beta_{11}^*$ (i.e., with regard to a z -transformed covariate $Z^* = (Z - E(Z))/\sqrt{\text{Var}(Z)}$). These are derived as follows:

$$\begin{aligned} \log[E(Y|X, Z^*)] &= \beta_{00}^* + \beta_{10}^* X + \beta_{01} Z^* + \beta_{11}^* X Z^* \\ &= \beta_{00}^* + \beta_{10}^* X + \beta_{01} \left(\frac{Z - E(Z)}{\sqrt{\text{Var}(Z)}} \right) + \beta_{11}^* X \left(\frac{Z - E(Z)}{\sqrt{\text{Var}(Z)}} \right) \\ &= \left(\beta_{00}^* - \frac{\beta_{01}^*}{\sqrt{\text{Var}(Z)}} E(Z) \right) + \left(\beta_{10}^* - \frac{\beta_{11}^*}{\sqrt{\text{Var}(Z)}} E(Z) \right) X \\ &\quad + \left(\frac{\beta_{01}^*}{\sqrt{\text{Var}(Z)}} \right) Z + \left(\frac{\beta_{11}^*}{\sqrt{\text{Var}(Z)}} \right) X Z \end{aligned}$$

Starting from the regression on X and the standardized Z^* , we derived the corresponding regression coefficients for the unstandardized Z , which are

$$\begin{aligned} \beta_{00} &= \beta_{00}^* - \frac{\beta_{01}^*}{\sqrt{\text{Var}(Z)}} E(Z), \\ \beta_{01} &= \frac{\beta_{01}^*}{\sqrt{\text{Var}(Z)}}, \\ \beta_{11} &= \frac{\beta_{11}^*}{\sqrt{\text{Var}(Z)}} \end{aligned}$$

The standardized regression coefficients are divided by the standard deviation of Z , because the covariates are on a different scale depending on their distribution in the simulation. We manipulated the effect size of the ATE, and thus the remaining regression coefficient β_{10} was not a free parameter but determined by

$$\beta_{10} = \log \left(\frac{\Delta \cdot \sqrt{\text{Var}(Y|X=0)} + M_Z(\beta_{01}) \exp(\beta_{00})}{\exp(\beta_{00}) M_Z(\beta_{01} + \beta_{11})} \right)$$

The standardized regression coefficients were chosen $\beta_{00}^* = 0$ and $\beta_{01}^* = 0.5$ to simulate a moderate covariate effect in the control group $X = 0$. The standardized interaction parameter was varied from no interaction ($\beta_{11}^* = 0$) to strong interaction ($\beta_{11}^* = 1$) with a step width of 0.25. As the derivations above suggest, the difference between the fixed-covariate and stochastic-covariate-based standard errors should increase with the strength of the interaction parameter. The population distributions of Z were chosen to resemble distinct distributional shapes: a standard normal distribution $\mathcal{N}(0, 1)$ for a symmetric shape with zero skewness and kurtosis, a Poisson distribution $\mathcal{P}(1)$ for a positively skewed shape, and a uniform distribution $U(-\sqrt{3}, \sqrt{3})$ for a flat shape with negative kurtosis. The parameters of $\mathcal{N}(0, 1)$ and $U(-\sqrt{3}, \sqrt{3})$ represent standardized variables, that is, $E(Z) = 0$, $\text{Var}(Z) = 1$. For the Poisson distribution, however, the parameter $\lambda = 1$ was chosen to create a moderate skew. We further varied the effect size Δ (Glass, 1976) of the ATE and the sample size N . In total, we evaluated 225 experimental scenarios (see Table 3 for more details on the design parameters).

The simulation consisted of three steps, which were carried out using the statistical software R (R Core Team, 2018). In the first step, N covariate values z_i were generated by drawing from the populational distribution of Z . In the second step, the values of the dichotomous treatment x_i were drawn from a Binomial distribution $B(1, p)$ where

$$p = P(X|Z^*) = 1 - \frac{1}{1 + \exp(Z^*)}$$

which was generated as a logistic function with standardized Z^* (i.e., the higher the value of Z , the higher the probability of getting into the treatment group). This means that we simulated a quasi-experimental setting, where treatment assignment is not random but depending on Z . As mentioned above, ATE and UTE are not equivalent in this case, and consideration of the covariate Z is necessary to estimate the average (causal) effect. In the third step, the conditional expectations $E(Y|X = x_i, Z = z_i)$ were computed and the observed y_i were drawn from a negative binomial distribution. The overdispersion parameter was $\alpha = 1$. Each experimental condition was replicated $R = 5000$ times. For more details and the corresponding R code, see the supplementary material.

Moment-Based Estimators by Distribution

For normal distributed $Z \sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$ with $\hat{\mu} = \bar{z}$ and $\hat{\sigma} = \hat{\sigma}_z$

$$\begin{aligned} \widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) &= \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \exp \left[\hat{\mu} (\hat{\beta}_{01} + \hat{\beta}_{11}) + \frac{\hat{\sigma}^2 (\hat{\beta}_{01} + \hat{\beta}_{11})^2}{2} \right] \\ &\quad - \exp(\hat{\beta}_{00}) \exp \left[\hat{\mu} \hat{\beta}_{01} + \frac{\hat{\sigma}^2 \hat{\beta}_{01}^2}{2} \right] \end{aligned}$$

For Poisson distributed $Z \sim \mathcal{P}(\hat{\lambda})$ where $\hat{\lambda} = \bar{z}$

$$\begin{aligned} \widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) = & \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \exp\left(\hat{\lambda} \left[\exp(\hat{\beta}_{01} + \hat{\beta}_{11}) - 1\right]\right) \\ & - \exp(\hat{\beta}_{00}) \exp\left(\hat{\lambda} \left[\exp(\hat{\beta}_{01}) - 1\right]\right) \end{aligned}$$

For Uniform distributed $Z \sim \text{U}(\hat{a}, \hat{b})$ where $\hat{a} = \bar{z} - \sqrt{3}\hat{\sigma}_z$ and $\hat{b} = \bar{z} + \sqrt{3}\hat{\sigma}_z$

$$\begin{aligned} \widehat{\text{ATE}}_{\text{MOM}}(\hat{\beta}, \hat{\theta}) = & \exp(\hat{\beta}_{00} + \hat{\beta}_{10}) \frac{\exp\left[\left(\hat{\beta}_{01} + \hat{\beta}_{11}\right)\hat{b}\right] - \exp\left[\left(\hat{\beta}_{01} + \hat{\beta}_{11}\right)\hat{a}\right]}{\left(\hat{\beta}_{01} + \hat{\beta}_{11}\right)(\hat{b} - \hat{a})} \\ & - \exp(\hat{\beta}_{00}) \frac{\exp\left[\hat{\beta}_{01}\hat{b}\right] - \exp\left[\hat{\beta}_{01}\hat{a}\right]}{\hat{\beta}_{01}(\hat{b} - \hat{a})} \end{aligned}$$

Estimation of Sample Moments for Moment-Based Approach

In our simulation, we used a two-step procedure for obtaining sample moment estimates and their covariance. The first step was based on a decomposition of the unconditional expectation and variance of Z with respect to the treatment groups:

$$\begin{aligned} E(Z) &= E(Z|X = 1) \cdot P(X = 1) + E(Z|X = 0) \cdot P(X = 0) \\ \text{Var}(Z) &= \text{Var}(Z|X = 1) \cdot P(X = 1) + \text{Var}(Z|X = 0) \cdot P(X = 0) \\ &\quad + (E(Z|X = 1) - E(Z))^2 \cdot P(X = 1) + (E(Z|X = 0) - E(Z))^2 \cdot P(X = 0) \end{aligned}$$

We used a multigroup structural equation model with stochastic group sizes to estimate the $X = x$ -conditional expectations and variances, and the probabilities $P(X = 0)$ and $P(X = 1)$.

$$\begin{aligned} \bar{z}_0 &= \hat{E}(Z|X = 0) \\ \bar{z}_1 &= \hat{E}(Z|X = 1) \\ s_0^2 &= \widehat{\text{Var}}(Z|X = 0) \\ s_1^2 &= \widehat{\text{Var}}(Z|X = 1) \\ p_0 &= \hat{P}(X = 0) \\ p_1 &= \hat{P}(X = 1) \end{aligned}$$

The sample mean $\bar{z}$ and variance s^2 were then obtained by

$$\begin{aligned} \bar{z} &= \bar{z}_0 \cdot p_0 + \bar{z}_1 \cdot p_1 \\ s^2 &= s_0^2 \cdot p_0 + s_1^2 \cdot p_1 + (\bar{z}_0 - \bar{z})^2 \cdot p_0 + (\bar{z}_1 - \bar{z})^2 \cdot p_1 \end{aligned}$$

This approach allows us to model the group sizes as random variables. Thus, variability in sample moment estimates due to different group weights are accounted for. For more

information on the underlying model and assumptions, see [Mayer et al. \(2016, p. 4\)](#).

In the second step, we used the estimated sample moments $\bar{z}$ and s^2 to compute the estimated parameters of the distribution. Depending on the distribution this was

$$\text{Normal } \mathcal{N}(\mu, \sigma): \quad \hat{\mu} = \bar{z}, \quad \hat{\sigma}^2 = s^2$$

$$\text{Poisson } \mathcal{P}(\lambda): \quad \hat{\lambda} = \bar{z}$$

$$\text{Uniform } U(a, b): \quad \hat{a} = \bar{z} - \sqrt{3s^2}, \quad \hat{b} = \bar{z} + \sqrt{3s^2}$$

For the three distributions in our simulation study, this procedure provided unbiased and efficient estimates for the parameters of the distribution. We especially compared the two-step procedure to conventional maximum likelihood estimates. For the uniformly distributed Z , the MLE provided biased estimates in small samples, while the two step-procedure was unbiased. For normally or Poisson distributed Z , point estimates from MLE and the two-step procedure were identical, however, the MLE was a little less efficient.

Data Generation Function

```

1  data.sim <- function(N, cohensd, b101, distribution){
2  # Standardized Regression Coefficients in Control Group
3  b000 <- 0      # Intercept
4  b001 <- 0.5    # Slope
5
6  # Remaining parameters are set depending on the covariate's distribution
7  if (distribution == "normal"){
8  muZ <- 0      # Mean of Z
9  sigmaZ <- 1   # Variance of Z
10 a000 <- b000-b001*muZ/sqrt(sigmaZ) # Unstandardized Intercept in Control Group
11 a001 <- b001/sqrt(sigmaZ) # Unstandardized Slope in Control Group
12 a101 <- b101/sqrt(sigmaZ) # Unstandardized Interaction Coefficient
13 varY0 <- 1.089 # Var(Y/X=0)
14 a100 <- log(sqrt(varY0)*cohensd + exp(.5*b001^2)) - .5*(b001 + b101)^2 #
    Unstandardized Coefficient of Treatment Variable
15 z <- rnorm(N, muZ, sigmaZ) # Z from a Standard Normal Distribution
16 } else if (distribution == "poisson"){
17 lambdaZ = 1 # Mean and Variance of Z
18 varY0 <- 1.06 # Var(Y/X=0)
19 a000 <- b000-b001*lambdaZ/sqrt(lambdaZ) # Unstandardized Intercept in Control Group
20 a001 <- b001/sqrt(lambdaZ) # Unstandardized Slope in Control Group
21 a101 <- b101/sqrt(lambdaZ) # Unstandardized Interaction Coefficient
22 a100 <- log(sqrt(varY0)*cohensd + exp(a000)*exp(lambdaZ*(exp(a001) - 1))) - log(exp(a000)
    )*exp(lambdaZ*(exp(a001+a101)-1))) # Unstandardized Coefficient of Treatment
    Variable
23 z <- rpois(N, lambdaZ) # Z from a Poisson distribution
24 } else if (distribution == "uniform"){
25 aZ <- -sqrt(3) # lower limit of Z
26 bZ <- sqrt(3) # upper limit of Z
27 muZ <- (sqrt(3)-sqrt(3))/2 # Mean of Z
28 sigmaZ <- sqrt((sqrt(3)+sqrt(3))^2/12) # Variance of Z
29 a000 <- b000-b001*muZ/sqrt(sigmaZ) # Unstandardized Intercept in Control Group
30 a001 <- b001/sqrt(sigmaZ) # Unstandardized Slope in Control Group
31 a101 <- b101/sqrt(sigmaZ) # Unstandardized Interaction Coefficient
32 varY0 <- 1.09 # Var(Y/X=0)
33 a100 <- log(sqrt(varY0)*cohensd + exp(b000)*(exp(b001*bZ) - exp(b001*aZ))/(b001*(bZ-aZ))
    ) -
34 log(exp(b000)*(exp((b001+b101)*bZ)-exp((b001+b101)*aZ))/((b001+b101)*(bZ-aZ))) #
    Unstandardized Coefficient of Treatment Variable
35 z <- runif(N, aZ, bZ) # Z from a Uniform distribution

```

```

36 }
37 zS <- scale(z)      # z-transformation of Z for generation of X
38 x <- rbinom(N, 1, plogis(zS))  # Treatment with treatment probability as logistic
    function of Z
39 muY <- exp(a000 + a001*z + a100*x + a101*x*z)  # Conditional Expectation of Y depending
    on Z and X
40 y <- rnbinom(N, mu = muY, size =1)  # Y from negative binomial distribution
41
42 d <- data.frame(y,x,z)
43 return(d)
44 }

```

Estimator Functions

```

1  library(lavaan)
2  library(car)
3
4  ##### function estimating TEMZ #####
5  computeTemz <- function(m1, d){
6    # Get regression coefficients
7    coefs <- coef(m1)
8    b00 <- coefs[1]
9    b10 <- coefs[2]
10   b01 <- coefs[3]
11   b11 <- coefs[4]
12
13   # Get mean of covariate z
14   zM <- mean(d$z)
15
16   # Compute and return TEMZ
17   temz <- exp(b00+b10+b01*zM+b11*zM) - exp(b00 + b01*zM)
18   return(temz)
19 }
20
21
22 ##### function estimating ATE (empirical approach, fixed covariate) #####
23 computeAveEmp <- function(m1, d){
24
25   # Compute all conditional effects
26   newdata0 <- newdata1 <- d
27   newdata0$x <- 0
28   newdata1$x <- 1
29
30   tau0 <- predict(m1, newdata=newdata0, type="response") ## tau0 is almost constant
31   tau1 <- predict(m1, newdata=newdata1, type="response") ## more variation in tau1
32   delta10 <- tau1 - tau0
33
34   # Compute VCOV of all conditional effects
35   vcovs <- vcov(m1)
36   mat <- cbind(delta10, tau1, d$z*delta10, d$z*tau1)
37   vcov_delta10 <- mat %*% vcovs %*% t(mat)
38
39   # Compute standard error for ATE
40   mat <- rbind(1/nrow(d),
41     as.numeric(d$x==0)/sum(d$x==0),
42     as.numeric(d$x==1)/sum(d$x))
43   vcov_eff <- mat %*% vcov_delta10 %*% t(mat)
44   se_Ave <- sqrt(vcov_eff[1,1])
45
46   # Return results

```

```

47   res <- list(Ave=mean(delta10),
48   se_Ave=se_Ave)
49
50   return(res)
51 }
52
53 ##### function estimating ATE (moment-based approach) #####
54 computeAveMom <- function(m1, d, distribution, se = "stochastic"){
55   # Get regression coefficients and there covariance
56   coefs <- coef(m1)
57   vcovs <- vcov(m1)
58
59   pnames <- c("g000", "g100", "g001", "g101")
60   names(coefs) <- row.names(vcovs) <- colnames(vcovs) <- pnames
61
62   # lavaan model for joint estimation of sample moments
63   modelz <- '
64   z ~ c(meanz001,meanz101)*1
65   z ~~ c(varz001,varz101)*z
66   group % c(groupw0,groupw1)*w
67   gw0 := groupw0
68   gw1 := groupw1
69   mz001 := meanz001
70   mz101 := meanz101
71   vz001 := varz001
72   vz101 := varz101
73   N := exp(gw0) + exp(gw1)
74   relfreq0 := exp(gw0)/N
75   relfreq1 := exp(gw1)/N
76   Ez1 := mz001*relfreq0 + mz101*relfreq1
77   Vz1 := vz001*relfreq0 + vz101*relfreq1 + relfreq0*(mz001-Ez1)^2 + relfreq1*(mz101-Ez1)^2
78   Px0 := relfreq0
79   Px1 := relfreq1
80   Pk0gx0 := relfreq0/Px0
81   Pk0gx1 := relfreq1/Px1
82   Ez1gx0 := mz001*Pk0gx0
83   Ez1gx1 := mz101*Pk0gx1
84   Vz1gx0 := vz001*Pk0gx0
85   Vz1gx1 := vz101*Pk0gx1
86   '
87   mz <- sem(modelz, data=d, group="x", group.label=c(0,1), group.w.free=TRUE)
88
89   ## augment coefs and vcovs
90   acoefs <- c(coefs, coef(mz, type="user")[-c(1:6)])
91
92   if (se == "stochastic"){
93     # for stochastic covariate
94     avcovs <- lav_matrix_bdiag(vcovs, lavInspect(mz, "vcov.def", add.class = FALSE))
95   } else if (se == "fixed"){
96     # for fixed covariate: covariance of distribution parameters fixed to zero
97     avcovs <- lav_matrix_bdiag(vcovs, lavInspect(mz, "vcov.def", add.class = FALSE)*0)
98   }
99
100   row.names(avcovs) <- colnames(avcovs) <- names(acoefts)
101
102   if (distribution == "normal"){
103     Eg1 <- deltaMethod(acoefts, "exp(g000+g100)*exp((g001+g101)*Ez1+((g001+g101)^2*Vz1/2))-
      exp(g000)*exp(g001*Ez1+(g001^2*Vz1/2))", avcovs, func="Eg1")
104   } else if (distribution == "uniform"){
105     Eg1 <- deltaMethod(acoefts,

```

```

106  "exp(g000+g100)*(exp((g001+g101)*(Ez1+sqrt(3*Vz1)))-exp((g001+g101)*(Ez1-sqrt(3*Vz1))))/
      ((g001+g101)*(2*sqrt(3*Vz1)))-exp(g000)*(exp(g001*(Ez1+sqrt(3*Vz1)))-exp(g001*(Ez1-
      sqrt(3*Vz1))))/(g001*(2*sqrt(3*Vz1)))",
107  avcovs, func="Eg1")
108  } else if (distribution == "poisson"){
109  Eg1 <- deltaMethod(acoeffs, "exp(g000+g100)*exp(Ez1*(exp(g001+g101)-1))-exp(g000)*exp(Ez1
      *(exp(g001)-1))", avcovs, func="Eg1")
110  } else if (distribution == "chisquare"){
111  Eg1 <- deltaMethod(acoeffs, "(exp(g000+g100)/(1-2*(g001+g101))^(Ez1/2))-(exp(g000)/(1-2*
      g001)^(Ez1/2))", avcovs, func="Eg1")
112  }
113
114  res <- list(Ave=Eg1$Estimate,
115  se_Ave=Eg1$SE)
116
117  return(res)
118
119  }

```

Simulation Procedure

```

1  library(doParallel)
2  library(foreach)
3  library(MASS)
4  source("dataSimulation.R")
5  source("estimators.R")
6
7  # Set design parameters
8  sampleN <- c(50, 100, 250, 500, 1000)
9  interact <- c(0, 0.25, 0.5, 0.75, 1)
10 cohensd <- c(-.5, 0, .5)
11 distribution <- c("normal", "poisson", "uniform")
12 reps <- 5000
13 conditions <- expand.grid(sampleN, interact, cohensd, distribution)
14 colnames(conditions) <- c("sampleN", "interact", "cohensd", "distribution")
15
16 ### Simulation procedure
17 #setup parallel backend to use 4 processors
18 cluster<-makeCluster(4, outfile = "debug.txt")
19 registerDoParallel(cluster)
20 results <- list()
21 for (i in 1:nrow(conditions)){
22  # Parallel replications within each scenario
23  res <- foreach::foreach(icount(reps), .combine = rbind, .packages = c("car", "lavaan", "
      MASS"), .errorhandling = "remove") %dopar% {
24    # Generate data set
25    d <- data.sim(conditions$sampleN[i],
26    conditions$cohensd[i],
27    conditions$interact[i],
28    conditions$distribution[i])
29    # Estimate NB-GLM
30    m1 <- glm.nb(y ~ x*z, data=d)
31    # Get results from ATE estimators
32    tmp <- c(unlist(computeAveEmp(m1, d, se = "fixed")),
33    unlist(computeAveAnaStoch(m1, d, distribution = conditions$distribution[i])),
34    unlist(computeAveAnaFixed(m1, d, distribution = conditions$distribution[i])),
35    unlist(computeAveAnaStoch(m1, d, distribution = "normal")),
36    computeTemz(m1, d),
37    m1$converged)
38    tmp

```

```
39 }
40 # Combine and save results
41 results <- data.frame(res,
42   conditions$sampleN[i],
43   conditions$cohensd[i],
44   conditions$interact[i],
45   conditions$distribution[i])
46 names(results) <- c("aveEmpFixed", "seEmpFixed",
47   "aveAnaStochastic", "seAnaStochastic",
48   "aveAnaFixed", "seAnaFixed",
49   "aveStochNorm", "seStochNorm",
50   "TEMZ",
51   "conv", "N", "cohensd", "interact", "distribution")
52 saveRDS(results, file = paste0("simulation/res_",
53   conditions$sampleN[i], "_",
54   conditions$cohensd[i], "_",
55   conditions$interact[i], "_",
56   conditions$distribution[i],
57   ".rds"))
58 }
59 stopCluster(cluster)
```

2.D Supplementary Material: CountEffects Package

As a supplement to this article, we provide an R package ([R Core Team, 2018](#)) called CountEffects, which comprises the functions related to the moment-based approach presented in this article.

It is available from GitHub and can be installed as follows:

```
1 library(devtools)
2 install_github('chkiefer/CountEffects')
```

3 Accounting for Latent Covariates

Abstract

The effectiveness of a treatment on a count outcome can be assessed using a negative binomial regression, where treatment effects are defined as the difference between the expected outcome under treatment and under control. These treatment effects can to date only be estimated if all covariates are manifest (observed) variables. However, some covariates are latent variables that are measured by multiple fallible indicators. In such cases, it is important to control for measurement error of covariates in order to avoid attenuation bias and to get unbiased treatment effect estimates. In this paper, we propose a new approach to compute average and conditional treatment effects in regression models with a logarithmic link function involving multiple latent and manifest covariates. We extend the approach presented by [Kiefer and Mayer \(2019\)](#) in several aspects: Building on a multigroup SEM framework for count variables instead of the generalized linear model, we allow for latent covariates and multiple covariates. We provide an illustrative example to explain the application and estimation in structural equation modeling software.

3.1 Introduction

The evaluation of treatment effects on count outcomes using randomized experiments is common in the social and health sciences. Covariates are often included in the analysis; these covariates may include latent psychological constructs that are not directly observable. In the field of drug and substance use prevention, the effects of interventions on the outcome variables daily/weekly count of alcoholic drinks or cigarettes consumed have been examined (Garrett et al., 2018; Neighbors et al., 2015) taking the covariates of readiness and social identity into account. Cognitive psychologists are often interested in how an intervention affects the counts of correctly-answered items on a cognitive test (Jobe et al., 2001; Nusser & Weinert, 2017), controlling for baseline mental abilities. In medicine, the effectiveness with which the count of MRI lesions and relapses in multiple sclerosis patients can be reduced is examined (Cascione et al., 2018; Cree et al., 2018), controlling for neurological health. Further examples are counts of risky sexual behavior (Hittner et al., 2016), accounting for drug use, count of students' antisocial behavior (Mazerolle et al., 2018), accounting for willingness to go to school, and days of absenteeism (Schaumborg & Flynn, 2017), accounting for guilt proneness.

These not directly observable covariates are usually measured with error. In psychology, for example, a participant's depressive state cannot be directly observed, but is measured by multiple fallible indicators (e.g., a symptom questionnaire). In such cases, the multiple indicators are assumed to measure a common factor or latent variable (Bentler, 1980; Bollen, 1989). Structural equation modeling offers a variety of options to specify a measurement model that allows a decomposition of true scores and measurement error, permitting the inclusion of the (error-free) latent variable in regression models (for an overview, see Bollen, 1989, for alternative approaches to deal with measurement error, see Schennach, 2016).

Commonly, manifest test scores are used for fallible covariates and the negative binomial (NB) regression model (Hilbe, 2007) is applied to investigate effects on count outcomes. NB regression is a generalized linear model (GLM) with logarithmic link function and NB random component. As in any nonlinear regression model, the coefficients of a NB regression do not correspond directly to average effectiveness on the original count scale of the outcome (Greene, 2007, p. 729) and can be difficult to interpret in presence of interaction effects. An alternative interpretation is in terms of marginal (Hilbe, 2007, p. 125) or treatment (Steyer & Nagel, 2017, p. 415) effects, which are defined as the differences between the conditional expectations under treatment and under control for each observation. The mean of these marginal effects is called average marginal/treatment effect (Greene, 2007, p. 730) and is computed over the marginal distribution of the covariates.

Typically, the reasons to adjust for baseline covariates in randomized experiments are twofold. First, the estimation precision for the (average) treatment effects can be increased compared to a simple difference-in-means estimate (Karrison & Kocherginsky, 2018; Miratrix, Sekhon, & Yu, 2012) – although it has previously been pointed out that

improved precision is not guaranteed (e.g., [Lin, 2013](#)). However, the covariates can help to explain variance in the outcome and thereby to reduce uncertainty. Second, by modeling treatment-covariate interactions, differential treatment effects can be examined. That is, the effect of a treatment is assumed to be moderated by the baseline covariates, resulting in distinct treatment effects for different persons. Thus, controlling for baseline covariates can help identifying who benefits most from a treatment.

Controlling for measurement error when examining covariate-adjusted treatment effects can be crucial. An error-prone covariate causes attenuation bias in estimation of the regression coefficients. That is, the slope coefficient of the error-prone covariate is typically attenuated towards zero. A similar phenomenon may occur, when factor scores are used. In this case, the values of the fallible covariate are not measured, but estimated with error ([Devlieger, Mayer, & Rosseel, 2016](#)). In (non-linear) models with several fallible covariates the magnitude and direction of attenuation is not easy to determine, as "the observed and underlying trends may be entirely different" ([Carroll, Ruppert, Stefanski, & Crainiceanu, 2010](#), p. 64).¹ As a consequence, attenuation bias can seriously impact the two aforementioned reasons for covariate-adjustment. First, a fallible covariate may explain less variance in the outcome variable and therefore the gain in precision can be lost. Second, magnitude and direction of regression coefficients may be altered, hindering examination of important and significant moderating variables.

While controlling for error-free covariates seems desirable in many cases, the common computation of marginal effects in NB regressions requires observed values of the covariate. That is, for each observed covariate value the expected count under treatment and under control is estimated. Then, the individual differences for these two expected counts are averaged over the empirical distribution. However, the values of a latent variable and their empirical distribution are unobserved and therefore cannot be used to compute the conditional expectations. Recently, [Kiefer and Mayer \(2019\)](#) proposed a moment-based approach to compute average treatment effects in regression models with a logarithmic link function that does not require observed covariate values. Instead, the moments of the covariates' marginal distribution are used to analytically compute the average treatment effect. Although this approach was only introduced for manifest variables, it can be adapted for latent variables. In this article, we extend the moment-based approach in two ways. First, we allow for latent covariates by using a multigroup structural equation modeling framework for count outcomes instead of the generalized linear model. This way, regression coefficients as well as treatment effects can be estimated based on an error-free covariate. Second, although the original approach only handles a single covariate, our extension allows for multiple latent and/or manifest covariates.

The manuscript is structured as follows: First, we introduce our terminology and definitions for average and conditional treatment effects. Second, the multigroup structural

¹In the supplementary material, we provide a Shiny app, that allows readers examine the effects of measurement error on regression coefficients in a Poisson regression with two (fallible) covariates. The computations are based on [Kukush, Schneeweis, and Wolf \(2004\)](#)

equation model with stochastic group sizes and a logarithmic link function is introduced as a basis for the extended approach. Third, we show how treatment effects can be estimated using moments of the covariates' distribution. Finally, we provide an illustrative example, in which the effects of a cognitive training intervention for elderly persons on their performance in inductive reasoning tasks are evaluated in a randomized experiment. We will demonstrate how the treatment effects can be estimated controlling for three latent covariates (i.e., depression, internal, and external locus of control).

3.2 Definition and Terminology of Treatment Effects

We start by introducing definitions for average and conditional effects of a binary treatment variable X with values $X = 0$ (control group) and $X = 1$ (treatment group) on a count outcome Y , taking into account a $(q + 1) \times 1$ -vector of (latent) covariates $\boldsymbol{\eta}$. Throughout this article, we focus on randomized experiments, and we follow the definitions and notation for treatment effects given by [Mayer et al. \(2016\)](#) and the stochastic theory of causal effects ([Steyer et al., 2014](#)). These authors define treatment effects as differences between the conditional expectations under treatment and under control. From a practical perspective, it is necessary that participants comply with their randomly assigned treatment and do not have missing values on the outcome Y , to ensure a causal interpretation of the effects. Further, one person's outcome must not depend on another person's treatment assignment and there must only be one level/version of the treatment (stable unit treatment value assumption, SUTVA; [Rubin, 1974](#)). All assumptions related to causality conditions, practical prerequisites for causal interpretation, and within the statistical model are summarized in Table 3.1. For NB regressions it is also common to use the (incidence) rate ratio as an effect size measure. At the end of this section, we will give a short excursus on how rate ratios are defined, how they relate to effects defined as differences, and why they can be difficult to interpret from a causal perspective.

3.2.1 Conditional Effects and the CE Function

A (conditional) treatment effect is defined as the difference between the expected outcomes under treatment and under control, that is:

$$CE(\boldsymbol{\eta}) = E(Y|X = 1, \boldsymbol{\eta}) - E(Y|X = 0, \boldsymbol{\eta}) \quad (3.1)$$

The conditional effect function $CE(\boldsymbol{\eta})$ gives a conditional treatment effect for any value of the covariates $\boldsymbol{\eta}$. The conditional expectations $E(Y|X = x, \boldsymbol{\eta})$ reflect the expected outcome given a treatment condition. For example, $E(Y|X = 1, \boldsymbol{\eta} = \mathbf{0})$ is the expected count of Y when a person with covariate value $\boldsymbol{\eta} = \mathbf{0}$ receives the treatment. The average effects discussed in this manuscript are aggregates of the conditional effect function. It is crucial that the CE function be causally unbiased. This is the case, when all potential confounders are independent from the treatment variable (i.e., treatment is randomly

Table 3.1: Assumptions within the extended moment-based approach

Assumptions	Explanation
<i>Causality conditions:</i>	
Unbiasedness of the effect function	Throughout this paper, we assume that participants are randomly assigned to either the control group ($X = 0$) or the treatment group ($X = 1$). This ensures $X \perp\!\!\!\perp \boldsymbol{\eta}$, independence of treatment and baseline covariates. However, there are other causality conditions to ensure unbiasedness of the effect function. For an overview, see Steyer et al. (2014) .
<i>Practical prerequisites for causal interpretation:</i>	
SUTVA	The Stable Unit Treatment Value Assumption consists of two parts (cf. Imbens & Rubin, 2015): 1. The potential outcome of one person must not depend on the treatment assignment of another person. 2. There must only be one level/version of the treatment.
Treatment compliance	Participants comply with their assigned treatment, that is, they actually participate in the assigned treatment condition
Fully observed data	The observed variables are fully observed; there are no missing data.
<i>Statistical model:</i>	
$Y \sim \text{NB}(\mu_Y, \phi_x)$	Outcome Y is $(X, \boldsymbol{\eta})$ -conditionally negative binomial distributed with mean μ_Y and variance $\mu_Y + \phi_x \mu_Y^2$
$\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\eta};x}, \boldsymbol{\Sigma}_{\boldsymbol{\eta};x})$	Latent covariates are X -conditionally multivariate normal distributed with mean vector $\boldsymbol{\mu}_{\boldsymbol{\eta};x}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta};x}$
$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Theta}_x)$	Observed indicators are $(X, \boldsymbol{\eta})$ -conditionally multivariate normally distributed with mean vector $(\boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta})$ and covariance matrix $\boldsymbol{\Theta}_x$
$N_0 \sim B(N, p_0), N_1 \sim B(N, 1 - p_0)$	Group sizes N_0 and N_1 are binomially distributed from a total sample size N with probability p_0
Treatment variable X , count outcome variable Y , latent covariates $\boldsymbol{\eta}$, and observed indicators $\mathbf{z}$. $\mu_Y = \exp(\boldsymbol{\beta}'_x \boldsymbol{\eta})$, $p_0 = P(X = 0) = \exp(\kappa)/(\exp(\kappa) + 1)$	

assigned). However, there are other causality conditions that ensure unbiasedness of the CE function when randomization does not hold (see Table 3.1). Further, the functional form of the effect function must be correctly specified. As we will see below, the functional form is determined by the regression model chosen for the outcome.

3.2.2 Average Effect and Effects at the Mean

We define the average effect of treatment $X = 1$ versus $X = 0$ as the unconditional expectation of the effect function $CE(\boldsymbol{\eta})$:

$$\begin{aligned} ATE &= E[CE(\boldsymbol{\eta})] \\ &= E[E(Y|X = 1, \boldsymbol{\eta}) - E(Y|X = 0, \boldsymbol{\eta})] \\ &= E[E(Y|X = 1, \boldsymbol{\eta})] - E[E(Y|X = 0, \boldsymbol{\eta})] \end{aligned} \tag{3.2}$$

The average treatment effect (ATE) is the effect in the entire population and is defined as the unconditional expectation of the conditional effects over the joint distribution of the covariates. In econometric literature, the ATE is also called the average marginal effect, because it is computed over the marginal distribution of the covariate. In health research, the term recycled predictions is also a common term for this kind of effect. If independence of X and $\boldsymbol{\eta}$ holds, the average treatment effect is also given by a simple mean difference between $X = 1$ and $X = 0$:

$$\begin{aligned} ATE &= E[E(Y|X = 1, \boldsymbol{\eta})] - E[E(Y|X = 0, \boldsymbol{\eta})] \\ &= E(Y|X = 1) - E(Y|X = 0) \end{aligned}$$

In randomized experiments, the group-mean difference and the average treatment effect are generally equivalent and causally unbiased (Steyer et al., 2014). When $\boldsymbol{\eta}$ is accounting for considerable variance in the outcome Y , the estimation of the ATE can be more precise compared with the simple mean difference (Karrison & Kocherginsky, 2018; Lin, 2013).

A related, but distinct, estimand, is the conditional effect at the mean of the covariates (CEM), that is, $CE(\boldsymbol{\mu}_{\boldsymbol{\eta}})$, where $\boldsymbol{\mu}_{\boldsymbol{\eta}}$ is the mean vector of $\boldsymbol{\eta}$. The CEM describes the conditional effect of the treatment for a person with average values on all covariates. In contrast, the ATE represents the average over all conditional effects. In non-linear models, the CEM and ATE are generally not equivalent. In a simulation study, Kiefer and Mayer (2019) showed that the CEM is often substantively smaller than the ATE for count outcomes. In linear models, however, the CEM and ATE are equivalent.

3.2.3 Excursus: Effects Defined as Ratio

In regression models with a log link, it is quite common to examine the incidence rate ratio (IRR) as a measure of (average) effectiveness, which can be denoted as

$$\text{IRR} = \frac{E(Y|X = 1, \boldsymbol{\eta} = \boldsymbol{\eta})}{E(Y|X = 0, \boldsymbol{\eta} = \boldsymbol{\eta})}.$$

The IRR is the ratio of conditional expectations under treatment and under control at a vector of covariate values $\boldsymbol{\eta}$. Usually, researchers look at $\boldsymbol{\eta} = \mathbf{0}$, as this is what the exponentiated regression coefficient of the treatment variable X represents. From a causal perspective, the IRR can be difficult to interpret for two reasons. First, in regression models without an interaction term, the IRR is constant across all values of the covariate. Thus, the IRR evaluated at the zero vector is the same as for every other vector of covariate values. In regression models with an interaction term, however, the IRR is a nonlinear function of the covariate values $\boldsymbol{\eta}$. This is shown in Figure 1. An interaction term is commonly added when differential treatment effects are of interest. If the covariates are mean-centered, the IRR at $\boldsymbol{\eta} = \mathbf{0}$ yields the IRR at the mean, which must not be confused with an average IRR_ϕ :

$$\text{IRR}_\phi = E \left[\frac{E(Y|X = 1, \boldsymbol{\eta})}{E(Y|X = 0, \boldsymbol{\eta})} \right].$$

Similar to effects based on differences, the average IRR and the IRR at the mean are generally not equivalent in a non-linear model with an interaction term. Moreover, neither of them is necessarily equivalent to the ratio of group-means

$$E \left[\frac{E(Y|X = 1, \boldsymbol{\eta})}{E(Y|X = 0, \boldsymbol{\eta})} \right] \neq \frac{E(Y|X = 1, \boldsymbol{\eta} = \boldsymbol{\eta})}{E(Y|X = 0, \boldsymbol{\eta} = \boldsymbol{\eta})} \neq \frac{E(Y|X = 1)}{E(Y|X = 0)},$$

not even in a randomized experiment. In conclusion, the interpretation of an IRR is restricted to a specific value of the covariate and does not necessarily reflect an average risk ratio if interaction effects are included. Second, both the IRR at the mean and the average IRR_ϕ (i.e., computing the IRR for each observation in the sample and computing an average) are non-collapsible. That is, even in a randomized controlled trial the IRR can substantively change if covariates are added to the model (Kiefer & Mayer, 2019), making it hard to identify the *true* causal effect. Non-collapsibility is a common issue in non-linear models. For an illustration of non-collapsibility for logistic regression, see Robinson and Jewell (1991); for a general overview, see Greenland et al. (1999). Thus, in models with an interaction term, an IRR might not provide a reliable measure of average effectiveness so that the ATE should be preferred, because it is collapsible (i.e., invariant with respect to additional covariates in randomized controlled trials) and an easily interpretable aggregate of individual effects.

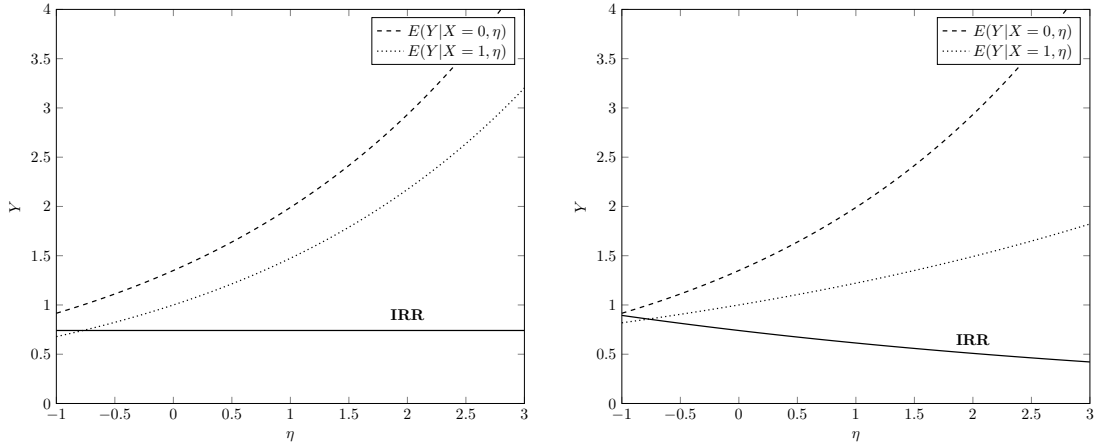


Figure 3.1

Left: Regression with logarithmic link function without interaction effect. The IRR is defined as the ratio of the conditional expectations at every $\eta = \eta$. If there is no interaction effect, the IRR is constant. Right: Regression with logarithmic link function with interaction effect. The IRR is a non-linear function $\exp(\beta_{10} + \beta_{11}\eta) / \exp(\beta_{00} + \beta_{01}\eta)$.

3.3 Negative Binomial Multigroup SEM (NB-MG-SEM)

The statistical framework for our approach is provided by a multigroup structural equation model with a logarithmic link function and stochastic group sizes. It allows us to estimate the measurement model for the latent covariates and negative binomial regression containing these latent covariates simultaneously. We chose a negative binomial regression over the well-known Poisson regression model, as it has a more flexible mean structure and allows for overdispersion. However, in scenarios without overdispersion, NB regression is equivalent to Poisson regression. First, we will briefly describe the different parts of the NB-MG-SEM and how they are involved in effect estimation. Second, we introduce the necessary assumptions to ensure unbiased treatment effect estimates. The maximum-likelihood estimation of the NB-MG-SEM is presented in Appendix 3.A.

The complete model consists of a group-specific structural model specifying the relation between the count outcome Y and the covariates $\boldsymbol{\eta}$ using a linear predictor, a logarithmic link function, and a conditional negative binomial distribution. The measurement structure is represented by a group-invariant linear measurement model relating a vector of manifest indicator variables $\mathbf{z} = (Z_1, Z_2, \dots)'$ to a vector of latent covariates $\boldsymbol{\eta} = (1, \eta_1, \dots, \eta_q)'$ (and a part for the group sizes):

$$\mu_Y = \exp(\boldsymbol{\beta}'_x \boldsymbol{\eta}) \quad (3.3)$$

$$\mathbf{z} = \boldsymbol{\nu} + \boldsymbol{\Lambda} \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad (3.4)$$

$$\log\left(\frac{n_0}{n_1}\right) = \kappa. \quad (3.5)$$

In Equations (3.3) to (3.5), Y has a negative binomial distribution with mean μ_Y and variance $\mu_Y + \phi_x \mu_Y^2$. ϕ_x is a group-specific overdispersion parameter; $\beta_x = (\beta_{x0}, \beta_{x1}, \dots, \beta_{xq})'$ is a vector of group-specific regression coefficients; ν is a group-invariant vector of measurement intercepts; Λ is a group-invariant matrix of factor loadings; ε is a vector of measurement error variables with mean vector $\mathbf{0}$ and covariance matrix Θ_x in group x ; and κ is a parameter for the log-transformed ratio of expected group frequencies. The probabilities for each group are $P(X = 0) = \exp(\kappa)/(\exp(\kappa) + 1)$ and $P(X = 1) = 1 - P(X = 0)$. The **linear** measurement model implies the mean vectors $\mu_{\eta;x}$ and covariance matrices $\Sigma_{\eta;x}$ of the latent covariates. Regression coefficients β_x , overdispersion parameters ϕ_x , measurement error variances Θ_x , mean vectors $\mu_{\eta;x}$, and covariance matrices $\Sigma_{\eta;x}$ of the latent covariates are group-specific as indicated by the x in the subscript. Effect estimation is based on mean vectors $\mu_{\eta;x}$ and covariance matrices $\Sigma_{\eta;x}$ of the latent covariates, the group-specific regression coefficients $\beta_x = (\beta_{x0}, \beta_{x1}, \dots, \beta_{xq})'$, and κ .

In the previous section, we assumed proper randomization of treatment assignment to define average and conditional treatment effects. Along with the NB-MG-SEM, we now introduce further assumptions to ensure unbiased estimation of the average treatment effects. We consider a sample of N independent and identically distributed observations $(Y, X, \mathbf{z})$. From a statistical perspective, we need to specify the distributions of the latent variables and the corresponding observed indicators. In maximum-likelihood estimation of structural equation models, it is common to choose the multivariate normal distribution. Thus, we assume $\eta \sim \mathcal{N}(\mu_{\eta;x}, \Sigma_{\eta;x})$ and $\mathbf{z}|\eta = \eta \sim \mathcal{N}(\nu + \Lambda\eta, \Theta_x)$, that is, the latent variables and the measurement errors are multivariate normally distributed. All assumptions made within the statistical model are summarized in Table 3.1. In the concluding section, we will address violations of these assumptions and how they can be addressed in applied settings.

The sample size requirements of the NB-MG-SEM depend on many things, for example, number of covariates and size of their slopes and interactions respectively, size of average effect, and desired power. As there is no comprehensive simulation study studying the power of the NB-MG-SEM approach, we recommend that researchers conduct a power analysis for their specific randomized experiment. In the supplemental material, we provide a Mplus template file for a Monte Carlo simulation of the NB-MG-SEM with a single latent covariate with three manifest indicators.

This framework of a multigroup SEM with linear predictor, a logarithmic link function, a conditionally negative binomial distributed dependent variable, and stochastic group sizes will be used to estimate average and conditional effects and to test hypotheses about these effects. Average and conditional effects and their standard errors will be computed according to parameter estimates and their variance-covariance matrix using the multivariate delta method. This model as presented here with a multinomial model for group sizes can be specified with the latent variable modeling software Mplus (Muthén & Muthén, 1998-2015) via the KNOWNCLASS option.

3.4 Extended Moment-Based Approach

In this section, we will show how conditional and average treatment effects can be computed based on the NB-MG-SEM. First, we will show how the conditional effect function and the average treatment effect can be defined using the count regression part of the NB-MG-SEM model. Second, we will give a brief introduction to the moment-based approach of [Kiefer and Mayer \(2019\)](#), providing an illustration of how it can provide closed-form solutions for average treatment effect computation. Third, the moment-based approach will be extended to multiple, latent covariates using the NB-MG-SEM framework.

3.4.1 Conditional Effects and CE Function

In the negative binomial multigroup SEM, the group-specific conditional expectation of Y is defined by

$$E(Y|X = x, \boldsymbol{\eta}) = \mu_Y = \exp(\boldsymbol{\beta}'_x \boldsymbol{\eta}).$$

Inserting this definition in the conditional effect function from Equation (3.1) yields

$$\begin{aligned} CE(\boldsymbol{\eta}) &= E(Y|X = 1, \boldsymbol{\eta}) - E(Y|X = 0, \boldsymbol{\eta}) \\ &= \exp(\boldsymbol{\beta}'_1 \boldsymbol{\eta}) - \exp(\boldsymbol{\beta}'_0 \boldsymbol{\eta}) \end{aligned} \tag{3.6}$$

The conditional expectation of Y is predicted by the negative binomial regression on the latent covariates $\boldsymbol{\eta}$. Thus, the CE function is the difference between the NB regression under treatment and under control.

3.4.2 Average Effect and Effect at the Mean

Above, we defined the average treatment effect as the expectation of the CE function. Consequently, for the NB-MG-SEM we can now insert Equation (3.6) into Equation (3.2):

$$\begin{aligned} ATE &= E[\exp(\boldsymbol{\beta}'_1 \boldsymbol{\eta}) - \exp(\boldsymbol{\beta}'_0 \boldsymbol{\eta})] \\ &= E[\exp(\boldsymbol{\beta}'_1 \boldsymbol{\eta})] - E[\exp(\boldsymbol{\beta}'_0 \boldsymbol{\eta})] \end{aligned}$$

The ATE can be computed via the unconditional expectation of the NB regression in the treatment group and in the control group. However, it can be difficult to estimate the ATE in practical settings because the difference of unconditional expectations represents a difference of two improper integrals. The moment-based approach of [Kiefer and Mayer \(2019\)](#) addresses this issue by replacing the improper integrals with closed-form expressions. In their paper, they suggested that for a single observed covariate Z the unconditional expectations of the NB regressions could be substituted with

$$M_Z(\beta_{x1}) = E[\exp(\beta_{x1} Z)],$$

the moment-generating function $M_Z(\beta_{x1})$ of Z . Based on the NB-MG-SEM, we can extend their approach for multiple (latent) covariates. In the NB-MG-SEM, $\boldsymbol{\eta}$ is assumed to be $X = x$ -conditional multivariate normally distributed (i.e., covariates are normally distributed given a value x of X) and therefore, the corresponding conditional moment-generating function is

$$M_{\boldsymbol{\eta}|X=x}(\boldsymbol{\beta}) = \exp \left(\boldsymbol{\mu}'_{\boldsymbol{\eta};x} \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\boldsymbol{\eta};x} \boldsymbol{\beta} \right).$$

As the latent covariates $\boldsymbol{\eta}$ are assumed to follow a X -conditional multivariate normal distribution, the unconditional distribution of the covariates $\boldsymbol{\eta}$ is a Gaussian mixture distribution and the corresponding unconditional moment-generating function is

$$\begin{aligned} M_{\boldsymbol{\eta}}(\boldsymbol{\beta}) &= \sum_{x=0}^1 P(X = x) M_{\boldsymbol{\eta}|X=x}(\boldsymbol{\beta}) \\ &= \sum_{x=0}^1 P(X = x) \exp \left(\boldsymbol{\mu}'_{\boldsymbol{\eta};x} \boldsymbol{\beta} + \frac{1}{2} \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\boldsymbol{\eta};x} \boldsymbol{\beta} \right) \end{aligned}$$

In conclusion, the average treatment effect can be expressed as:

$$ATE = \sum_{x=0}^1 P(X = x) \cdot \left[\exp \left(\boldsymbol{\mu}'_{\boldsymbol{\eta};x} \boldsymbol{\beta}_1 + \frac{1}{2} \boldsymbol{\beta}'_1 \boldsymbol{\Sigma}_{\boldsymbol{\eta};x} \boldsymbol{\beta}_1 \right) - \exp \left(\boldsymbol{\mu}'_{\boldsymbol{\eta};x} \boldsymbol{\beta}_0 + \frac{1}{2} \boldsymbol{\beta}'_0 \boldsymbol{\Sigma}_{\boldsymbol{\eta};x} \boldsymbol{\beta}_0 \right) \right]. \quad (3.7)$$

The computation of the ATE in these expressions is based on regression coefficients $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$, mean vectors $\boldsymbol{\mu}_{\boldsymbol{\eta};0}, \boldsymbol{\mu}_{\boldsymbol{\eta};1}$, and covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\eta};0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta};1}$ of the covariates $\boldsymbol{\eta}$. All these parameters are readily available in the previously presented negative binomial multigroup structural equation model.

In contrast, the conditional effect at the mean does not include the covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\eta};0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta};1}$ of the covariates $\boldsymbol{\eta}$ and is given by

$$CE(\boldsymbol{\mu}_{\boldsymbol{\eta}}) = \exp(\boldsymbol{\mu}_{\boldsymbol{\eta}}' \boldsymbol{\beta}_1) - \exp(\boldsymbol{\mu}_{\boldsymbol{\eta}}' \boldsymbol{\beta}_0),$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}} = \sum_{x=0}^1 P(X = x) \cdot \boldsymbol{\mu}_{\boldsymbol{\eta};x}$.

In conclusion, we extended the moment-based approach by [Kiefer and Mayer \(2019\)](#) in two ways. First, we use a NB-MG-SEM instead of a generalized linear model in order to include latent covariates to the analysis. Second, we allow for using the moment-based approach with multiple covariates $\boldsymbol{\eta}$ by using a multivariate moment-generating function.

3.5 Illustrative Example

To illustrate the use of the extended moment-based approach, we use data from the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) study ([Ball et al.](#),

2002; Jobe et al., 2001). The ACTIVE study is a large randomized controlled trial designed to examine the effectiveness of cognitive interventions with older adults. In our analysis, we examine the effectiveness of the ACTIVE-reasoning training ($X = 1$, $N = 699$) compared to the no-contact control group ($X = 0$, $N = 698$). The training program consisted of ten, 60-75 minute sessions in which participants learned and practiced **solving** problems that require linear thinking and that follow a serial pattern (Jobe et al., 2001). Accordingly, we investigate posttest performance in an inductive reasoning assessment (i.e. *word series*, 30 items, 6-minutes) as an outcome Y . Word series evaluate how well an individual can follow a series of words in logical sequence by providing the appropriate word that follows the last word in a list (e.g., January, January, February, March, March, April, ...). Thus, Y reflects the count of correctly-answered items.

As stated above, in applied settings, some assumptions of our model may not fully hold. In this example, there were cases with missing data in observed baseline variables or the outcome variable. Participants with missing data on all variables ($N = 117$) have been excluded from analysis. In addition, we used a full information maximum likelihood approach to account for the remaining missing data.² The final sample size for our analysis was $N = 1280$. There was also some treatment noncompliance, as only 89% of the participants fully completed training. In the discussion, we will discuss how noncompliance can be addressed in applied settings. In our illustrative example, we report an intention-to-treat analysis (Hollis & Campbell, 1999).

3.5.1 Measures

We controlled for potential interactions between treatment and depression (η_1), treatment and external locus of control (η_2), and treatment and internal locus of control (η_3). Depressive symptoms were assessed using the 12-item version of the Center of Epidemiological Studies-Depression (CESD-12) scale (Radloff, 1977). Participants rated the frequency of several depressive symptoms (e.g., feeling sad) during the last week on a 4-point scale from 0 = *never* to 3 = *5-7 days*. The *internal* and *external* locus of control (LoC) dimensions were assessed using a 36-item version of the Personality in Intellectual Aging Contexts (PIC) Inventory Control scale (Lachman, Baltes, Nesselroade, & Willis, 1982). We used the powerful others subscale to measure external LoC. Lachman et al. (1982, p. 491) defined the powerful others scale by a "dependence and reliance on other people for accomplishing intellectual tasks, due to the belief that others are better able to carry out such tasks" (e.g., I cannot figure out sale prices of items unless someone helps me). The internal LoC scale was defined by the belief that the "responsibility for modifications or maintenance of intellectual functioning lies within one's own control" (e.g., If I studied a map carefully, I could figure out how to get around in a strange place). These items were rated on a 6-point scale from 0 = *strongly disagree* to 5 = *strongly agree*. Depression and

²Analogous to the standard maximum likelihood estimation described in Appendix 3.A, the full information maximum likelihood approach used in this example (i.e., its implementation on Mplus) treats the count outcome as conditionally negative binomial distributed.

locus of control have been linked to the effectiveness and outcomes of cognitive training in previous studies (e.g., Lohman et al., 2012; Parisi et al., 2014; Wolinsky et al., 2009; Zahodne et al., 2015).

In our example, all three covariates η_1 , η_2 , and η_3 are latent. The baseline depression score η_1 was measured with three parcels (i.e., sum score of four items each; Z_{11}, Z_{12}, Z_{13}). The baseline external LoC η_2 was represented with two parcels (Z_{21}, Z_{22} ; each parcel is mean score of three items). The baseline internal locus of control η_3 was also represented by two parcels (Z_{31}, Z_{32} ; each parcel is mean score of three items). We used random item parcels for simplicity in our illustrative example. For cautions about the use of parcels see Marsh, Lüdtke, Nagengast, Morin, and von Davier (2013) and Sterba and Rights (2016). For a comparison model with only manifest covariates, we also computed mean scores over the respective parcels for all three covariates. The reliabilities were $\alpha = .80$ for depression, $\alpha = .77$ for the external locus of control scale, and $\alpha = .67$ for the internal locus of control scale.

3.5.2 Model

The basis for our effect analysis was a two-group structural equation model with group-invariant linear measurement models, and a structural model with linear predictor, a logarithmic link function, and a negative binomial distribution for the dependent variable. The measurement models were chosen to be τ -congeneric for all three covariates, as we had no previous assumptions about the true scores (e.g., τ -equivalence). All three latent variables were scaled to their respective first indicator variable, for example, a 1-unit change on η_1 corresponds to a 1-unit change on Z_{11} . Although intercepts ν and loadings Λ were group-invariant, the measurement error variances of ε were allowed to vary between groups. The measurement model is expressed as:

$$\begin{pmatrix} Z_{11} \\ Z_{12} \\ Z_{13} \\ Z_{21} \\ Z_{22} \\ Z_{31} \\ Z_{32} \end{pmatrix} = \begin{pmatrix} 0 \\ \nu_{12} \\ \nu_{13} \\ 0 \\ \nu_{22} \\ 0 \\ \nu_{32} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{12} & 0 & 0 \\ \lambda_{13} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{22} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{32} \end{pmatrix} \cdot \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}.$$

The means, variances, and covariances of the latent variables were estimated. In our example, the group-specific mean vector $\mu_{\eta;x}$ and covariance matrix $\Sigma_{\eta;x}$ of the latent covariates η were:

$$\mu_{\eta;x} = \begin{pmatrix} \mu_{1;x} \\ \mu_{2;x} \\ \mu_{3;x} \end{pmatrix}, \quad \Sigma_{\eta;x} = \begin{pmatrix} \sigma_{11;x}^2 & \sigma_{12;x}^2 & \sigma_{13;x}^2 \\ \sigma_{12;x}^2 & \sigma_{22;x}^2 & \sigma_{23;x}^2 \\ \sigma_{13;x}^2 & \sigma_{23;x}^2 & \sigma_{33;x}^2 \end{pmatrix}.$$

In the structural model with linear predictor, a logarithmic link function and a negative binomial distribution for the dependent variable, the three latent covariates η_1 , η_2 , and η_3 were linked to the expected outcome of the count variable Y :

$$\begin{aligned} E(Y|X = 0, \eta_1, \eta_2, \eta_3) &= \exp(\beta_{00} + \beta_{01}\eta_1 + \beta_{02}\eta_2 + \beta_{03}\eta_3) \\ E(Y|X = 1, \eta_1, \eta_2, \eta_3) &= \exp(\beta_{10} + \beta_{11}\eta_1 + \beta_{12}\eta_2 + \beta_{13}\eta_3) \end{aligned}$$

The regression coefficients β were group-specific, that is, the effects of the covariates could be different between both groups. The treatment effects are not constant in such models due to the logarithmic link function and the interaction between the treatment and covariates. For a single covariate, this is illustrated in Figure 3.1. Even in the absence of a treatment-covariate interaction (Figure 3.1, on the left), the conditional effects – represented by the difference between the two regression lines – depend on the value of the covariates. This result differs from that in linear regression models. In presence of a treatment-covariate interaction (Figure 3.1, on the right), the CE function is also not a constant effect. For instance, cognitive training might be more effective for patients with a low baseline depression score than for a patients with a high baseline depression score. Models without interactions on the log scale can be estimated as special cases by constraining the corresponding regression coefficients to be equal, e.g., $\beta_{01} = \beta_{11}$. The count of correctly-answered items on Y was assumed to follow a conditional negative binomial distribution with overdispersion parameter ϕ_x . Finally, the model for group sizes of the training group (sample size n_1) and the control group (sample size n_0) is expressed as:

$$\log\left(\frac{n_0}{n_1}\right) = \kappa.$$

The full model is displayed in Figure 3.2. Note that the regression model with logarithmic link function is indicated by curved arrows in contrast to straight arrows for linear regressions.

3.5.3 Average and Conditional Treatment Effects

Using the parameters of the two-group model presented in the previous section we can compute the treatment effects. Based on Equation (3.7) the average treatment effect in

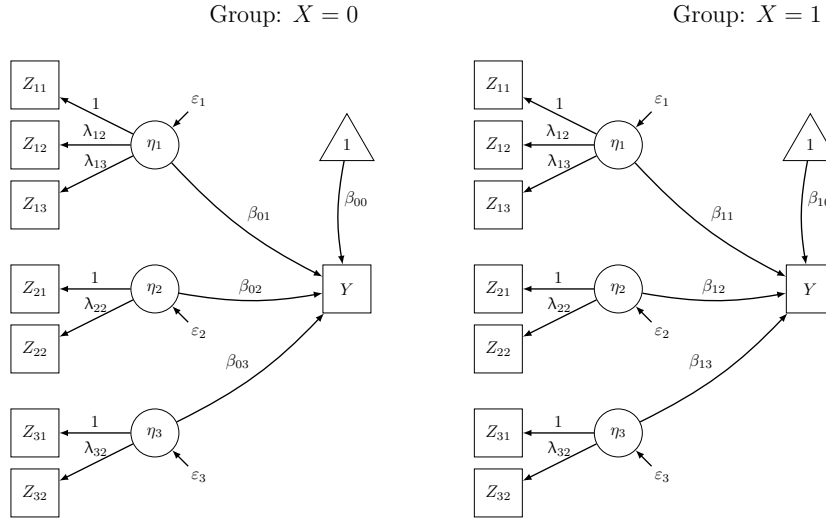


Figure 3.2

Path diagram for our illustrative example depicting the group invariant linear measurement models for the latent baseline depression η_1 , the latent baseline external LoC η_2 , and the latent baseline internal LoC η_3 , the structural model with linear predictor and logarithmic link function (indicated by curved arrows) of the posttest count of correctly-answered items Y on the latent baseline variables η_1 , η_2 , and η_3 in each of the two groups.

our concrete example is

$$\begin{aligned}
 ATE = & \sum_{x=0}^1 P(X = x) \exp \left[\beta_{10} + \beta_{11}\mu_{1;x} + \beta_{12}\mu_{2;x} + \beta_{13}\mu_{3;x} \right. \\
 & + \frac{1}{2} \left(\beta_{11}^2 \sigma_{11;x}^2 + \beta_{12}^2 \sigma_{22;x}^2 + \beta_{13}^2 \sigma_{33;x}^2 \right) \\
 & + \frac{1}{2} \left(\beta_{11}\beta_{12}\sigma_{12;x}^2 + \beta_{11}\beta_{13}\sigma_{13;x}^2 + \beta_{12}\beta_{13}\sigma_{23;x}^2 \right) \left. \right] \\
 & - \exp \left[\beta_{00} + \beta_{01}\mu_{1;x} + \beta_{02}\mu_{2;x} + \beta_{03}\mu_{3;x} \right. \\
 & + \frac{1}{2} \left(\beta_{01}^2 \sigma_{11;x}^2 + \beta_{02}^2 \sigma_{22;x}^2 + \beta_{03}^2 \sigma_{33;x}^2 \right) \\
 & + \frac{1}{2} \left(\beta_{01}\beta_{02}\sigma_{12;x}^2 + \beta_{01}\beta_{03}\sigma_{13;x}^2 + \beta_{02}\beta_{03}\sigma_{23;x}^2 \right) \left. \right],
 \end{aligned}$$

where $P(X = 0) = \exp(\kappa)/(\exp(\kappa) + 1)$ and $P(X = 1) = 1 - P(X = 0)$.

The analyses were carried out using *Mplus* 8.2 (Muthén & Muthén, 1998-2015). The *Mplus* syntax is given in the Supplementary Material.

3.5.4 Comparative Analyses

In order to compare the results from the extended moment-based approach, we also conducted the effect analysis using manifest (fallible) test scores S_1 (depression), S_2 (external LoC), and S_3 (internal LoC). We estimated a generalized linear model with a negative bi-

nomial random component (NB-GLM). The manifest test scores were computed as means over the respective parcels, thereby maintaining the same scale and 1-unit change interpretation of the coefficients as with the latent variables. The model equation for the corresponding NB-GLM is:

$$E(Y|X, S_1, S_2, S_3) = \exp(\gamma_{00} + \gamma_{10}X + \gamma_{01}S_1 + \gamma_{02}S_2 + \gamma_{03}S_3 + \gamma_{11}XS_1 + \gamma_{12}XS_2 + \gamma_{13}XS_3)$$

Note that we use γ as the notation for the coefficients in the NB-GLM; NB-GLM is usually estimated across groups. The group-specific coefficients in the β -notation can be transformed into the γ -coefficients. This is shown in Appendix 3.B. The average treatment effect was obtained by averaging over the observed values of the conditional effect function

$$\begin{aligned} ATE = \sum_{i=1}^N \exp(\gamma_{00} + \gamma_{10} + \gamma_{01}S_{1i} + \gamma_{02}S_{2i} + \gamma_{03}S_{3i} + \gamma_{11}S_{1i} + \gamma_{12}S_{2i} + \gamma_{13}S_{3i}) \\ - \exp(\gamma_{00} + \gamma_{01}S_{1i} + \gamma_{02}S_{2i} + \gamma_{03}S_{3i}) \end{aligned} \quad (3.8)$$

Standard errors for the average treatment effect can be computed using the delta method (for more details see [Greene, 2007](#)). As the delta method approximates asymptotically correct standard errors, its performance in finite samples is dependent on sample size. For a single covariate, [Kiefer and Mayer \(2019\)](#) showed that the delta method could provide adequate standard errors for the ATE even in small samples ($N = 50$), assuming the covariates were normally distributed and/or interaction effects were modest. However, for highly skewed covariates and large interaction effects, the standard errors were severely biased. Additionally, in very small samples the point estimates of the NB-GLM can be biased (e.g., $N = 50$, cf. [Lord, 2006](#)).

3.5.5 Results

Currently, it is not possible to evaluate model fit using a χ^2 -test or other fit statistics, as these are not provided for count outcome models by Mplus. In order to assess the goodness-of-fit of our measurement only model, we additionally estimated a multigroup measurement only-model for our three latent covariates (i.e., deleting the structural part). The measurement only-model indicated a good fit with $\chi^2(30) = 46.522$, RMSEA = .034, CFI = .993, TLI = .990, SRMR = .027. The parameters of the group-invariant measurement model are shown in Table 3.2. Note that this model fit applies to the measurement-only model and does not necessarily represent to the full NB-MG-SEM model. It provides us with the information that this component of the full model fits well.

Regression Coefficients

Results for the two-group structural equation model with logarithmic link function (NB-MG-SEM) are summarized in Table 3.3. We report the group-specific regression coefficients and the parameters of the X -conditional distributions of the covariates. In the

Table 3.2: Maximum likelihood estimates for the measurement model for the three latent covariates

Parameter	Estimate	SE	p	Parameter	Estimate	SE	p
$\hat{\nu}_{11}$	0.000			$\hat{\lambda}_{11}$	1.000		
$\hat{\nu}_{12}$	-0.090	0.111	.421	$\hat{\lambda}_{12}$	1.286	0.059	< .001
$\hat{\nu}_{13}$	-0.437	0.115	< .001	$\hat{\lambda}_{13}$	1.348	0.064	< .001
$\hat{\nu}_{21}$	0.000			$\hat{\lambda}_{21}$	1.000		
$\hat{\nu}_{22}$	-0.475	0.099	< .001	$\hat{\lambda}_{22}$	1.171	0.058	< .001
$\hat{\nu}_{31}$	0.000			$\hat{\lambda}_{31}$	1.000		
$\hat{\nu}_{32}$	1.579	0.224	< .001	$\hat{\lambda}_{32}$	0.695	0.054	< .001
$\widehat{Var}(\varepsilon_{11;0})$	1.556	0.136	< .001	$\widehat{Var}(\varepsilon_{11;1})$	1.367	0.163	< .001
$\widehat{Var}(\varepsilon_{12;0})$	1.401	0.169	< .001	$\widehat{Var}(\varepsilon_{12;1})$	1.516	0.162	< .001
$\widehat{Var}(\varepsilon_{13;0})$	1.391	0.177	< .001	$\widehat{Var}(\varepsilon_{13;1})$	1.219	0.190	< .001
$\widehat{Var}(\varepsilon_{21;0})$	0.807	0.083	< .001	$\widehat{Var}(\varepsilon_{21;1})$	0.641	0.065	< .001
$\widehat{Var}(\varepsilon_{22;0})$	0.400	0.074	< .001	$\widehat{Var}(\varepsilon_{22;1})$	0.452	0.057	< .001
$\widehat{Var}(\varepsilon_{31;0})$	0.387	0.055	< .001	$\widehat{Var}(\varepsilon_{31;1})$	0.406	0.059	< .001
$\widehat{Var}(\varepsilon_{32;0})$	0.300	0.045	< .001	$\widehat{Var}(\varepsilon_{32;1})$	0.317	0.037	< .001

SE = standard error; Intercepts ν , Factor Loadings λ , Residual variances $Var(\varepsilon)$

control group, the estimate for the intercept was $\hat{\beta}_{00} = 2.435$, $p < .001$, 95 % $CI = [1.794, 3.076]$, which means that for a person with scores of 0 on all three baseline covariates, the expected outcome is $\hat{E}(Y|X = 0, \eta_1 = \eta_2 = \eta_3 = 0) = \exp(2.435) = 11.416$ correctly solved items. The estimated regression coefficients of η_1 was $\hat{\beta}_{01} = -0.052$, $p = .001$, 95% $CI = [-0.081, -0.023]$. For a 1-unit increase in baseline depression η_1 , the expected count is multiplied by $\exp(-0.052) = 0.949$. So in the control group, each additional unit on the 4-point depression scale at baseline is associated with a 5 percent decrease in the expected count of correctly-answered items at posttest. Similarly, for the baseline external locus of control η_2 and baseline internal LoC η_3 the estimated regression coefficients were $\hat{\beta}_{02} = -0.207$, $p < .001$, 95% $CI = [-0.303, -0.111]$ and $\hat{\beta}_{03} = 0.088$, $p = .169$, 95% $CI = [-0.037, 0.213]$.

In the training group, the estimate for the intercept was $\hat{\beta}_{10} = 2.748$, $p < .001$, 95% $CI = [1.966, 3.530]$, which implies that for a person with scores of 0 on all three baseline covariates, the expected outcome is $\hat{E}(Y|X = 1, \eta_1 = \eta_2 = \eta_3 = 0) = \exp(2.748) = 15.611$ correctly-answered items. The estimated regression coefficients of η_1 was $\hat{\beta}_{11} = -0.005$, $p = .773$, 95% $CI = [-0.040, 0.030]$. For a 1-unit increase in baseline depression η_1 , the expected count is multiplied by $\exp(-0.005) = 0.995$. Thus, in the training group, each additional unit on the 4-point depression scale at baseline is associated with a *non-significant* 0.5 percent decrease in the expected count of correctly-answered items at posttest. Similarly, for the baseline external LoC η_2 and baseline internal LoC η_3 , the estimated regression coefficients were $\hat{\beta}_{12} = -0.218$, $p < .001$, 95% $CI = [-0.338, -0.098]$ and $\hat{\beta}_{13} = 0.049$, $p = .524$, 95% $CI = [-0.102, 0.200]$.

Table 3.3: Maximum likelihood estimates for the two-group structural equation model with stochastic group sizes and logarithmic link function

Parameter	Estimate	SE	p	Parameter	Estimate	SE	p
$X = 0$:				$X = 1$			
Regression Coefficients:							
$\hat{\beta}_{00}$	2.435	0.327	< .001	$\hat{\beta}_{10}$	2.748	0.399	< .001
$\hat{\beta}_{01}$	− 0.052	0.015	.001	$\hat{\beta}_{11}$	− 0.005	0.018	.773
$\hat{\beta}_{02}$	− 0.207	0.049	< .001	$\hat{\beta}_{12}$	− 0.218	0.061	< .001
$\hat{\beta}_{03}$	0.088	0.064	.169	$\hat{\beta}_{13}$	0.049	0.077	.524
Means:							
$\hat{\mu}_{1;0}$	1.652	0.080	< .001	$\hat{\mu}_{1;1}$	1.747	0.081	< .001
$\hat{\mu}_{2;0}$	1.643	0.055	< .001	$\hat{\mu}_{2;1}$	1.670	0.050	< .001
$\hat{\mu}_{3;0}$	3.885	0.044	< .001	$\hat{\mu}_{3;1}$	3.912	0.042	< .001
(Co-)Variances:							
$\hat{\sigma}_{11;0}$	1.831	0.256	< .001	$\hat{\sigma}_{11;1}$	2.000	0.240	< .001
$\hat{\sigma}_{22;0}$	0.864	0.098	< .001	$\hat{\sigma}_{22;1}$	0.761	0.075	< .001
$\hat{\sigma}_{33;0}$	0.554	0.072	< .001	$\hat{\sigma}_{33;1}$	0.464	0.062	< .001
$\hat{\sigma}_{12;0}$	0.452	0.086	< .001	$\hat{\sigma}_{12;1}$	0.561	0.078	< .001
$\hat{\sigma}_{13;0}$	− 0.344	0.075	< .001	$\hat{\sigma}_{13;1}$	− 0.417	0.070	< .001
$\hat{\sigma}_{23;0}$	− 0.531	0.058	< .001	$\hat{\sigma}_{23;1}$	− 0.487	0.044	< .001
Group Size:							
κ	0.016	0.056	.780				

SE = standard error; β_{x0} = Intercept in group x , β_{x1} = Slope for latent depression in group x , β_{x2} = Slope for latent external locus of control in group x , β_{x3} = Slope for latent internal locus of control in group x .

Table 3.4: Estimated average and conditional effects of therapy ($X = 1$ vs. $X = 0$) compared to the control group in the simulated data example.

Effect	Estimate	SE	p
$\widehat{ATE}$	2.589	0.287	< .001
$\widehat{CEM}$	2.607	0.282	< .001
Group Probability			
$P(X = 0)$	0.504	0.014	< .001
$P(X = 1)$	0.496	0.014	< .001

SE = standard error; ATE = Average Treatment Effect; CEM = Conditional Effect at Mean

Average and Conditional Treatment Effects

In this example, the estimated average treatment effect of reasoning training ($\widehat{ATE} = 2.589$, 95% CI [2.026, 3.152]) is statistical significant. It provides an estimate of the difference between expected counts in the treatment group and the control group adjusted for the covariates. The difference between $\hat{E}[\hat{E}(Y|X = 0, \eta_1, \eta_2, \eta_3)] = 10.774$ and $\hat{E}[\hat{E}(Y|X = 1, \eta_1, \eta_2, \eta_3)] = 13.362$ implies that the expected count of correctly-answered items increased by 2.589, on average, due to the reasoning training. The conditional effect at the mean ($\widehat{CEM} = 2.607$, 95% CI [2.054, 3.160]) is close to the ATE in this case. Results on average and conditional treatment effects are summarized in Table 3.4.

Comparative model

In the comparative analysis, we used listwise deletion for missing data, resulting in a sample of $N = 871$. The estimated regression coefficients and the estimated average treatment effect are given in Table 3.5. In addition, the corresponding parameter estimates from the NB-MG-SEM are shown. Note that β and γ coefficients are linear transformations of each other and we computed the corresponding γ coefficients for the NB-MG-SEM. The derivation and computation is shown in Appendix 3.B

Magnitude and direction of the regression coefficients is similar across both models, with exception of the slope coefficient γ_{03} , which is significant in the NB-GLM ($\gamma_{03;GLM} = 0.102, p < .001$), but not significant in the NB-MG-SEM ($\gamma_{03;SEM} = 0.088, p = .169$). Thus, both models could lead to different conclusion about the influence of internal LoC on cognitive skills. However, as the three score variables have near acceptable ($\alpha = .67$) to good ($\alpha = .80$) reliability, one would expect only small amounts of attenuation bias. For illustrative purposes, we also computed the comparative model with shorter scales for the covariates (by selecting the three (LoC) or four (depression) most reliable items for each scale) to investigate the attenuation bias for (partly) lower reliabilities ($\alpha = .56 - .81$). Depending on which covariates were replaced with a shortened scale, there were significant differences for other regression coefficients (i.e., the regression coefficient for the treatment γ_{10} , and the treatment-baseline depression interaction effect γ_{11}). These findings underline the complex effects of attenuation bias when multiple fallible covariates

Table 3.5: Comparative Analyses

Parameter	NB-GLM			NB-MG-SEM		
	Estimate	SE	p	Estimate	SE	p
γ_{00}	2.287	0.149	< .001	2.435	0.327	< .001
γ_{10}	0.104	0.105	.325	0.314	0.515	.543
γ_{01}	-0.045	0.012	< .001	-0.052	0.015	.001
γ_{02}	-0.145	0.021	< .001	-0.207	0.049	< .001
γ_{03}	0.102	0.031	.001	0.088	0.064	.169
γ_{11}	0.016	0.008	.036	0.047	0.024	.048
γ_{12}	0.010	0.015	.504	-0.011	0.079	.893
γ_{13}	-0.009	0.022	.678	-0.039	0.100	.695
ATE	2.660	0.323	< .001	2.589	0.287	< .001

SE = standard error; γ_{x0} = Intercept in group x , γ_{x1} = Slope/interaction for manifest depression, γ_{x2} = Slope/interaction for manifest external locus of control, γ_{x3} = Slope/interaction for manifest internal locus of control

are present.

Estimates and inferences for the average treatment effect are similar for both the moment-based approach and the empirical approach based on the NB-GLM. It has previously been shown that the empirical-distribution-based approach from Equation (3.8) produces similar results to the moment-based approach in absence of a strong interaction effect (Kiefer & Mayer, 2019), as is the case here.

3.6 Summary and Conclusions

In this paper, we presented and illustrated a new method of adjusting for latent covariates when estimating average and conditional treatment effects for count outcomes. We extended the moment-based approach of Kiefer and Mayer (2019) in two ways. First, a negative binomial multigroup structural equation model was used instead of the generalized linear model. The NB-MG-SEM offers the opportunity to estimate simultaneously the measurement model for the latent covariates and the negative binomial regression structural model based on these covariates. Thus, the model accounts for measurement error in the covariates and thereby minimizes attenuation bias and the consequent loss of power for tests of the coefficients in the regression model. Second, we allow for multiple manifest and/or latent covariates by using multivariate moment-generating functions for multivariate normal and Gaussian mixture distributions. This way an arbitrary number of continuous covariates can be included in the model. In non-randomized observational studies it can be crucial to account for several covariates. Finally, we provided an illustrative example to show how the extended moment-based approach can be applied to real data. In our example, the effect of a cognitive reasoning training group was compared to a no-contact control group on the count of correctly-answered items in a cognitive reasoning test. We considered depression, external, and internal locus of control as latent covariates. The NB-MG-SEM and the treatment effects were estimated in Mplus.

3.6.1 Limitations and Further Research

There are some aspects not covered in this article that could provide starting points for further research and extensions of the approach: Throughout this paper, the assumption of randomized treatment assignment allowed a causal interpretation of treatment effects. In nonrandomized observational studies these effects are not necessarily causal effects. However, theories of causality, such as Rubin's causal model (Imbens & Rubin, 2015; Rubin, 1974) or the stochastic theory of causal effects (Steyer et al., 2014), provide further causality conditions (besides randomization) under which causal effects are estimable. Usually, these conditions require a careful selection of covariates, which then are controlled with regression-adjustment (e.g., Mayer et al., 2016) or propensity score methods (Rosenbaum & Rubin, 1983). For an overview on design and analysis in quasi-experimental settings see also Reichhardt (2019). Sengewald et al. (2018) and Sengewald and Pohl (2019) examined under which conditions controlling for latent covariates (opposed to fallible covariates) can lead to causal effect estimates in observational studies. Note that in observational studies the summands of the ATE in Equation (3.7) can be interpreted as treatment effects given a treatment condition, as discussed by, for example, Geneletti and Dawid (2011).

Within the NB-MG-SEM and the effect estimation, we assume the covariates follow a conditional multivariate normal distribution. In some applications, this assumption may not be met. Our approach can be extended for cases with severe violations of the conditional normality assumption in two different ways: First, an alternative multivariate distribution with a known moment-generating function can be chosen. For example, the multivariate skew-normal distribution suggested by Azzalini and Valle (1996) can account for skewness in the covariates; this distribution has a closed form moment-generating function for effect estimation. In this case, the likelihood of the NB-MG-SEM needs to be adapted. Second, a Gaussian mixture distribution can be used to approximate several kinds of multivariate distributions. Mayer, Umbach, Flunger, and Kelava (2017) used a (Bayesian) approach for treatment effect estimation in a nonlinear regression with a skewed covariate, where the covariates distribution was approximated using a finite mixture of normal distributions (for more details see also Kelava & Nagengast, 2012; Kelava, Nagengast, & Brandt, 2014). As the moment-generating function for such Gaussian mixture distributions is known, this approach could be combined with our extended moment-based approach.

Sometimes, treatment assignment is performed at the cluster level, that is groups of individuals are randomly assigned to the treatment conditions. In this case, the assumption of identically and independently distributed observations can be violated. In order to account for dependencies between subjects from the same cluster and also to examine the effect contextual (i.e., cluster-level) covariates, a multilevel framework could be applied. Mayer, Nagengast, Fletcher, and Steyer (2014) proposed a multilevel multigroup SEM for regressions with identity link function, which could be adapted for regression models with a logarithmic link function.

In our illustrative example, about 11 % of subjects did not fully complete the cognitive training. This so-called treatment noncompliance can compromise the causal interpretation of a randomized experiment, if the decision to not comply is related to the treatment and/or the outcome. [Meier \(1991\)](#) provides applied scenarios, under which treatment effects are confounded by noncompliance. There are several techniques to deal with treatment noncompliance. We used an intention-to-treat approach, that is, subjects were analyzed as if they had fully completed training. The major advantage of this approach is that group comparability is maintained. However, it has been argued that intention-to-treat estimates confound treatment effects with treatment non-compliance. In cases in which the null hypothesis cannot be rejected, it remains unclear whether this result is due to a non-effective treatment or to treatment non-compliance ([Sheiner & Rubin, 1995](#)). There are several other techniques to deal with treatment noncompliance, some of which allow for integrating available information on noncompliance into the analysis. For a recent overview of these techniques see [Sagarin et al. \(2014\)](#).

An advantage of the NB-MG-SEM model over a NB-GLM, is that it allows for easy missing data handling by using full information maximum likelihood (FIML) estimation ([Arbuckle, 1996](#)). In our comparative analysis we used listwise deletion for the NB-GLM, which may be another reason for possible differences between both approaches. In more comprehensive analyses, multiple imputation provides an alternative for listwise deletion in the NB-GLM ([Enders, 2010](#)).

Appendix

3.A Estimation

Let us consider a count outcome variable Y , a binary treatment variable X , a set of K latent covariates $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$, and a set of L observed indicator variables $\mathbf{z} = (Z_1, \dots, Z_L)$. Further, we want to estimate a statistical model as presented in Equations (3.3), (3.4), and (3.5) using maximum-likelihood techniques. Assuming a sample of N independently and identically distributed observations, the complete data likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}|Y, X, \mathbf{z}, \boldsymbol{\eta}) = \prod_{i=1}^N f(y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}_i|\boldsymbol{\theta})$$

where $y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}_i$ are the i -th observed value of the aforementioned variables. As we do not have observed values $\boldsymbol{\eta}_i$, we actually have to optimize the observed data likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|Y, X, \mathbf{z}) &= \prod_{i=1}^N f(y_i, x_i, \mathbf{z}_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^N \int_{\boldsymbol{\eta}} f(y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}|\boldsymbol{\theta}) d\boldsymbol{\eta} \end{aligned}$$

Note that marginalizing over the joint density of the complete data is only one way to optimize this likelihood. It would also be possible to use an EM algorithm (Dempster, Laird, & Rubin, 1977) to solve the complete data likelihood, but this would have been beyond the scope of this manuscript.

The joint density $f(y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}_i|\boldsymbol{\theta})$ can be factorized in several conditional densities:

$$\begin{aligned} f(y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}_i|\boldsymbol{\theta}) &= P(X = x_i|\boldsymbol{\theta}) \cdot f(y_i, \mathbf{z}_i, \boldsymbol{\eta}_i|\boldsymbol{\theta}, x_i) \\ &= P(X = x_i|\boldsymbol{\theta}) \cdot f(\boldsymbol{\eta}_i|\boldsymbol{\theta}, x_i) \cdot f(y_i, \mathbf{z}_i|\boldsymbol{\theta}, x_i, \boldsymbol{\eta}_i) \\ &= P(X = x_i|\boldsymbol{\theta}) \cdot f(\boldsymbol{\eta}_i|\boldsymbol{\theta}, x_i) \cdot f(\mathbf{z}_i|\boldsymbol{\theta}, x_i, \boldsymbol{\eta}_i) \cdot f(y_i|\boldsymbol{\theta}, x_i, \boldsymbol{\eta}_i, \mathbf{z}_i) \\ &= P(X = x_i|\boldsymbol{\theta}_1) \cdot f(\boldsymbol{\eta}_i|\boldsymbol{\theta}_2, x_i) \cdot f(\mathbf{z}_i|\boldsymbol{\theta}_3, x_i, \boldsymbol{\eta}_i) \cdot f(y_i|\boldsymbol{\theta}_4, x_i, \boldsymbol{\eta}_i, \mathbf{z}_i) \end{aligned}$$

In the last step, we decomposed the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4)$ into four parts corresponding to the four conditional densities. The parameters to estimate in the respective densities are:

$$\begin{aligned} \boldsymbol{\theta}_1 &= \kappa \\ \boldsymbol{\theta}_2 &= (\boldsymbol{\mu}_{\boldsymbol{\eta};x}, \boldsymbol{\Sigma}_{\boldsymbol{\eta};x}) \\ \boldsymbol{\theta}_3 &= (\boldsymbol{\nu}, \boldsymbol{\Lambda}, \boldsymbol{\Theta}_x) \\ \boldsymbol{\theta}_4 &= (\boldsymbol{\phi}_x, \boldsymbol{\beta}_x) \end{aligned}$$

In the NB-MG-SEM, the conditional densities are defined by the model assumptions, which are summarized in Table 3.1. Here, $P(X = x_i|\boldsymbol{\theta}_1)$ reflect the group weights, where the group size n_0 (and n_1 respectively) is assumed to binomial distributed. The group-specific distribution of the latent variables is captured in $f(\boldsymbol{\eta}_i|\boldsymbol{\theta}_2, x_i)$, the group-specific and $\eta = \eta_i$ -conditional distribution of the observed indicators $\mathbf{z}$ is given by $f(\mathbf{z}_i|\boldsymbol{\theta}_3, x_i, \boldsymbol{\eta}_i)$ and the group-specific negative binomial regression of Y on $\boldsymbol{\eta}$ is given by $f(y_i|\boldsymbol{\theta}_4, x_i, \boldsymbol{\eta}_i, \mathbf{z}_i)$. Thus, the conditional densities in our case are

$$\begin{aligned} P(X = x_i|\boldsymbol{\theta}_1) &= (1 - x_i) \frac{\exp(\kappa)}{\exp(\kappa) + 1} + x_i \frac{1}{\exp(\kappa) + 1} \\ f(\boldsymbol{\eta}_i|\boldsymbol{\theta}_2, x_i) &= \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_x|}} \exp\left(-\frac{1}{2}(\boldsymbol{\eta}_i - \boldsymbol{\mu}_\eta)' \boldsymbol{\Sigma}_x^{-1} (\boldsymbol{\eta}_i - \boldsymbol{\mu}_\eta)\right) \\ f(\mathbf{z}_i|\boldsymbol{\theta}_3, x_i, \boldsymbol{\eta}_i) &= \frac{1}{\sqrt{(2\pi)^L |\boldsymbol{\Theta}_x|}} \exp\left(-\frac{1}{2}(\mathbf{z}_i - \boldsymbol{\eta}_i)' \boldsymbol{\Theta}_x^{-1} (\mathbf{z}_i - \boldsymbol{\eta}_i)\right) \\ f(y_i|\boldsymbol{\theta}_4, x_i, \boldsymbol{\eta}_i, \mathbf{z}_i) &= \exp\left[y_i \log\left(\frac{\phi_x \mu_{Y_i}}{1 + \phi_x \mu_{Y_i}}\right) - \frac{1}{\phi_x} \log(1 + \phi_x \mu_{Y_i})\right. \\ &\quad \left. + \log \Gamma(y_i + \frac{1}{\phi_x}) - \log \Gamma(y_i + 1) - \log \Gamma(\frac{1}{\phi_x})\right] \end{aligned}$$

where $\mu_{Y_i} = \exp(\boldsymbol{\beta}'_x \boldsymbol{\eta}_i)$

As the group weights do not depend on the latent covariates, we can simplify the observed data density as follows:

$$\begin{aligned} f(y_i, x_i, \mathbf{z}_i|\boldsymbol{\theta}) &= \int_{\boldsymbol{\eta}} f(y_i, x_i, \mathbf{z}_i, \boldsymbol{\eta}|\boldsymbol{\theta}) d\boldsymbol{\eta} \\ &= P(X = x_i|\boldsymbol{\theta}) \cdot \int_{\boldsymbol{\eta}} f(y_i, \mathbf{z}_i, \boldsymbol{\eta}|\boldsymbol{\theta}, x_i) d\boldsymbol{\eta} \end{aligned}$$

and hence, the corresponding log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta}|Y, X, \mathbf{z}) = \sum_{i=1}^N \log[P(X = x_i|\boldsymbol{\theta})] + \sum_{i=1}^N \log\left[\int_{\boldsymbol{\eta}} f(y_i, \mathbf{z}_i, \boldsymbol{\eta}|\boldsymbol{\theta}, x_i) d\boldsymbol{\eta}\right]$$

As there is no closed form solution for $\int_{\boldsymbol{\eta}} f(y_i, \mathbf{z}_i, \boldsymbol{\eta}|\boldsymbol{\theta}, x_i) d\boldsymbol{\eta}$, numerical integration is required.

3.B Regression Coefficients Transformation

For the two-group NB-MG-SEM, we chose a β -notation for the regression coefficients. In the generalized linear model with interaction term, we chose an γ -notation instead. In this section, we show that the coefficients are linear transformations of each other and can easily be computed from each other.

Let the conditional expectation of Y be

$$E(Y|X, \eta_1, \eta_2, \eta_3) = \exp(\gamma_{00} + \gamma_{10}X + \gamma_{01}\eta_1 + \gamma_{02}\eta_2 + \gamma_{03}\eta_3 + \gamma_{11}X\eta_1 + \gamma_{12}X\eta_2 + \gamma_{13}X\eta_3)$$

then, setting $X = 1$ and $X = 0$ respectively provides

$$E(Y|X = 0, \eta_1, \eta_2, \eta_3) = \exp(\gamma_{00} + \gamma_{01}\eta_1 + \gamma_{02}\eta_2 + \gamma_{03}\eta_3)$$

$$E(Y|X = 1, \eta_1, \eta_2, \eta_3) = \exp((\gamma_{00} + \gamma_{10}) + (\gamma_{01} + \gamma_{11})\eta_1 + (\gamma_{02} + \gamma_{12})\eta_2 + (\gamma_{03} + \gamma_{13})\eta_3)$$

As we can see, the β -coefficients correspond to the γ -coefficients in the following way:

$$\beta_{00} = \gamma_{00}$$

$$\beta_{01} = \gamma_{01}$$

$$\beta_{02} = \gamma_{02}$$

$$\beta_{03} = \gamma_{03}$$

$$\beta_{10} = \gamma_{00} + \gamma_{10}$$

$$\beta_{11} = \gamma_{01} + \gamma_{11}$$

$$\beta_{12} = \gamma_{02} + \gamma_{12}$$

$$\beta_{13} = \gamma_{03} + \gamma_{13}$$

and vice versa

$$\gamma_{00} = \beta_{00}$$

$$\gamma_{01} = \beta_{01}$$

$$\gamma_{02} = \beta_{02}$$

$$\gamma_{03} = \beta_{03}$$

$$\gamma_{10} = \beta_{10} - \beta_{00}$$

$$\gamma_{11} = \beta_{11} - \beta_{01}$$

$$\gamma_{12} = \beta_{12} - \beta_{02}$$

$$\gamma_{13} = \beta_{13} - \beta_{03}$$

3.C Supplementary Material: Shiny App on Attenuation Bias

```
1 library(shiny)
2 library(nleqslv)
3 attenuationFunc <- function(x, par){
4   # Fixed and true parameters
5   b0 <- par[1]
6   b1 <- par[2]
7   b2 <- par[3]
8   covZ <- par[4]
9   veps1 <- par[5]
10  veps2 <- par[6]
11
12  # Resulting, attenuated parameters
13  a0 <- x[1]
14  a1 <- x[2]
15  a2 <- x[3]
16
17  # set of equations
18  y <- numeric( 3 )
19  y[1] <- exp(b0 + .5*b1^2 + .5*b2^2 + .5*b1*b2*covZ) - exp(a0 + .5*a1^2*(veps1+1) + .5*a2
20    ^2*(veps2+1) + .5*a1*a2*covZ)
21  y[2] <- (b1 + .5*b2*covZ)*exp(b0 + .5*b1^2 + .5*b2^2 + .5*b1*b2*covZ) - (a1*(veps1+1) +
22    .5*a2*covZ)*exp(a0 + .5*a1^2*(veps1+1) + .5*a2^2*(veps2+1) + .5*a1*a2*covZ)
23  y[3] <- (b2 + .5*b1*covZ)*exp(b0 + .5*b1^2 + .5*b2^2 + .5*b1*b2*covZ) - (a2*(veps2+1) +
24    .5*a1*covZ)*exp(a0 + .5*a1^2*(veps1+1) + .5*a2^2*(veps2+1) + .5*a1*a2*covZ)
25  y
26 }
27
28 # Define UI for application that draws a histogram
29 ui <- fluidPage(
30
31   # Application title
32   titlePanel("Attenuation Bias in Poisson Regression"),
33
34   # Sidebar with a slider input for number of bins
35   sidebarLayout(
36     sidebarPanel(
37       sliderInput("b0",
38         "Regression parameter b0",
39         min = -1,
40         max = 1,
41         value = 0,
42         step = .1),
43       sliderInput("b1",
44         "Regression parameter b1",
45         min = -1,
46         max = 1,
47         value = 0,
48         step = .1),
49       sliderInput("b2",
50         "Regression parameter b2",
51         min = -1,
52         max = 1,
53         value = 0,
54         step = .1),
55       sliderInput("covZ",
56         "Correlation of latent variables",
57         min = -1,
58         max = 1,
```

```

56         value = 0,
57         step = .1),
58     sliderInput("veps1",
59         "Error variance in first variable",
60         min = 0,
61         max = 10,
62         value = 0,
63         step = .5),
64     sliderInput("veps2",
65         "Error variance in second variable",
66         min = 0,
67         max = 10,
68         value = 0,
69         step = .5)
70 ),
71
72 # Show a plot of the generated distribution
73 mainPanel(
74     fluidRow(
75         splitLayout(cellWidths = c("50%", "50%"), plotOutput("regPlot"), plotOutput("
76             regSPlot"))
77     ),
78     fluidRow(
79         splitLayout(cellWidths = c("50%", "50%"), uiOutput('regFormulaEta'), uiOutput('
80             regFormulaScore'))
81     )
82 )
83
84 # Define server logic required to draw a histogram
85 server <- function(input, output) {
86     output$regPlot <- renderPlot({
87         # generate bins based on input$bins from ui.R
88         b0 <- input$b0
89         b1 <- input$b1
90         b2 <- input$b2
91
92         x <- seq(-5, 5, length = 30)
93         y <- x
94         f1 <- function(x,y){exp(b0 + b1*x + b2*y)}
95         z <- outer(x, y, f1)
96         persp(x, y, z, theta = 60, phi = 0, col = "lightblue", xlab = "eta1", ylab = "eta2",zlab
97             = "Y")
98     })
99     output$regSPlot <- renderPlot({
100         # generate bins based on input$bins from ui.R
101         b0 <- input$b0
102         b1 <- input$b1
103         b2 <- input$b2
104         covZ <- input$covZ
105         veps1 <- input$veps1
106         veps2 <- input$veps2
107         par <- c(b0,b1,b2,covZ,veps1,veps2)
108
109         xstart <- c(0,0,0)
110         res <- nleqslv(xstart, attenuationFunc, par = par)
111         a0 <- res$x[1]
112         a1 <- res$x[2]
113         a2 <- res$x[3]

```

```

114
115 x <- seq(-5, 5, length = 30)
116 y <- x
117 f2 <- function(x,y){exp(a0 + a1*x + a2*y)}
118 z <- outer(x, y, f2)
119 persp(x, y, z, theta = 60, phi = 0, col = "lightblue", xlab = "S1", ylab = "S2", zlab =
    "y")
120
121 })
122
123 output$regFormulaEta <- renderUI({
124   b0 <- input$b0
125   b1 <- input$b1
126   b2 <- input$b2
127
128   withMathJax(sprintf("Regression under latent variable:
129   $$E(Y|\eta_{1},\eta_{2}) = \exp(0.2f + 0.2f \eta_{1} + 0.2f \eta_{2})$$", b0, b1,
    b2))
130 })
131
132 output$regFormulaScore <- renderUI({
133   b0 <- input$b0
134   b1 <- input$b1
135   b2 <- input$b2
136   covZ <- input$covZ
137   veps1 <- input$veps1
138   veps2 <- input$veps2
139   par <- c(b0,b1,b2,covZ,veps1,veps2)
140
141   xstart <- c(0,0,0)
142   res <- nleqslv(xstart, attenuationFunc, par = par)
143   a0 <- res$x[1]
144   a1 <- res$x[2]
145   a2 <- res$x[3]
146
147   withMathJax(sprintf("Regression under score variable:
148   $$E(Y|S_1,S_2) = \exp(0.2f + 0.2f S_1 + 0.2f S_2)$$", a0, a1, a2))
149 })
150 }
151
152 # Run the application
153 shinyApp(ui = ui, server = server)

```

3.D Supplementary Material: Mplus Syntax for Illustrative Example

Mplus syntax:

```

1  TITLE: Analysis ACTIVE Data
2  DATA:
3  FILE = "data.dat";
4
5  VARIABLE:
6  NAMES = dv treat z11 z12 z21 z22 z31 z32 z41 z42 z43;
7  USEVARIABLES dv treat z11 z12 z21 z22 z41 z42 z43;
8
9  ! Specify outcome variable "dv" as count with (conditional) negative binomial (nb)
   distribution
10 COUNT ARE dv (nb);
11
12 ! KNOWNCLASS option for multigroup model with stochastic group size
13 classes = cell( 2 );
14 knownclass = cell( treat=0 treat=2);
15 MISSING=.;
16
17 ANALYSIS:
18 ! for KNOWNCLASS option, TYPE = MIXTURE is required
19 TYPE = MIXTURE;
20
21 ! INTEGRATION for numerical integration of the likelihood on 8 cores of the CPU
22 ALGORITHM = INTEGRATION;
23 process = 8;
24
25 ! Specification of the full model
26 MODEL:
27 %overall%
28 ! Overall specification of the model
29
30 ! Measurement models for the latent covariates
31 depress BY z41 z42 z43; ! Measurement model for depression = eta1
32 power BY z11 z12; ! Measurement model for external LoC = eta2
33 internal BY z21 z22; ! Measurement model for internal LoC = eta3
34 [z41@0 z11@0 z21@0]; ! Mean of respective first indicator fixed to 0
35
36 ! Negative binomial regression E(Y|eta1, eta2, eta3)
37 dv ON depress power internal;
38
39 ! Group size model
40 [cell#1] (LC0);
41
42 %cell#1%
43 ! Group-specific parameters for the control group X=0
44
45 ! Negative binomial regression E(Y|X=0,eta1,eta2,eta3)
46 ! i.e., exp(beta00 + beta01*eta1 + beta02*eta2 + beta03*eta3)
47 dv ON power(beta0_2)
48 internal(beta0_3)
49 depress(beta0_1);
50 [dv] (beta0_0); ! Intercept
51 dv; ! Overdispersion parameter phi_0
52
53 ! Means, variances and covariances of the latent variables
54 ! Means

```

```

55 [depress] (Ez1C0); !E(eta1|X=0)
56 [power] (Ez2C0); !E(eta2|X=0)
57 [internal] (Ez3C0); !E(eta3|X=0)
58
59 !Variances
60 depress (Vz1C0); ! Var(eta1|X=0)
61 power (Vz2C0); ! Var(eta2|X=0)
62 internal (Vz3C0); ! Var(eta3|X=0)
63
64 !Covariances
65 power WITH depress(Vz12C0); ! Cov(eta3, eta1|X=0)
66 internal WITH depress(Vz13C0); ! Cov(eta2, eta1|X=0)
67 power WITH internal(Vz23C0); ! Cov(eta3, eta2|X=0)
68
69 ! Measurement error in observed indicators
70 z41 z42 z43 z11 z12 z21 z22; ! Var(epsilon_ij|X=0)
71
72
73 %cell#2%
74 ! Group-specific parameters for the treatment group X=1
75
76 ! Negative binomial regression E(Y|X=1,eta1,eta2,eta3)
77 ! i.e., exp(beta10 + beta11*eta1 + beta12*eta2 + beta13*eta3)
78 dv ON power(beta1_2)
79 internal(beta1_3)
80 depress(beta1_1);
81 [dv] (beta1_0); ! Intercept
82 dv; ! Overdispersion parameter phi_1
83
84 ! Means, variances and covariances of the latent variables
85 ! Means
86 [depress] (Ez1C1); ! E(eta1|X=1)
87 [power] (Ez2C1); ! E(eta2|X=1)
88 [internal] (Ez3C1); ! E(eta3|X=1)
89
90 !Variances
91 depress (Vz1C1); ! Var(eta1|X=1)
92 power (Vz2C1); ! Var(eta2|X=1)
93 internal (Vz3C1); ! Var(eta3|X=1)
94
95 !Covariances
96 power WITH depress(Vz12C1); ! Cov(eta3, eta1|X=1)
97 internal WITH depress(Vz13C1); ! Cov(eta2, eta1|X=1)
98 power WITH internal(Vz23C1); ! Cov(eta3, eta2|X=1)
99
100 ! Measurement error in observed indicators
101 z41 z42 z43 z11 z12 z21 z22; ! Var(epsilon_ij|X=1)
102
103 MODEL CONSTRAINT:
104 ! Computation of further parameters from the model parameters
105 NEW(PC0 PC1
106 Eg1x0
107 Eg1x1
108 Eg1
109 muEta1 muEta2 muEta3 CEM1
110 Md1 Md2 Md3
111 );
112
113 ! Group probability
114 PC0 = EXP(LC0) / (EXP(LC0) +1);
115 PC1 = 1 - PC0;

```

```

116
117 ! Effects for Group 0
118 Eg1x0 =
119 EXP(beta1_0 + beta1_1*Ez1C0 + beta1_2*Ez2C0 + beta1_3*Ez3C0 +
120 .5*(beta1_1^2*Vz1C0 + beta1_2^2*Vz2C0 + beta1_3^2*Vz3C0
121 + beta1_1*beta1_2*Vz12C0 + beta1_1*beta1_3*Vz13C0 + beta1_2*beta1_3*Vz23C0))-
122 EXP(beta0_0 + beta0_1*Ez1C0 + beta0_2*Ez2C0 + beta0_3*Ez3C0 +
123 .5*(beta0_1^2*Vz1C0 + beta0_2^2*Vz2C0 + beta0_3^2*Vz3C0
124 + beta0_1*beta0_2*Vz12C0 + beta0_1*beta0_3*Vz13C0 + beta0_2*beta0_3*Vz23C0));
125
126 ! Effects for Group 1
127 Eg1x1 =
128 EXP(beta1_0 + beta1_1*Ez1C1 + beta1_2*Ez2C1 + beta1_3*Ez3C1 +
129 .5*(beta1_1^2*Vz1C1 + beta1_2^2*Vz2C1 + beta1_3^2*Vz3C1
130 + beta1_1*beta1_2*Vz12C1 + beta1_1*beta1_3*Vz13C1 + beta1_2*beta1_3*Vz23C1))-
131 EXP(beta0_0 + beta0_1*Ez1C1 + beta0_2*Ez2C1 + beta0_3*Ez3C1 +
132 .5*(beta0_1^2*Vz1C1 + beta0_2^2*Vz2C1 + beta0_3^2*Vz3C1
133 + beta0_1*beta0_2*Vz12C1 + beta0_1*beta0_3*Vz13C1 + beta0_2*beta0_3*Vz23C1));
134
135 ! Average Effects
136 Eg1 = PC0*Eg1x0 + PC1*Eg1x1;
137
138 ! Conditional Effect at Mean
139 muEta1 = PC0*Ez1C0+PC1*Ez1C1;
140 muEta2 = PC0*Ez2C0+PC1*Ez2C1;
141 muEta3 = PC0*Ez3C0+PC1*Ez3C1;
142 CEM1 =
143 EXP(beta1_0 + beta1_1*muEta1 + beta1_2*muEta2 + beta1_3*muEta3)
144 -EXP(beta0_0 + beta0_1*muEta1 + beta0_2*muEta2 + beta0_3*muEta3);

```

4 Accounting for Non-Normal Covariates

Abstract

The effects of a treatment or an intervention on a count outcome is often of interest in applied research. When controlling for additional covariates, a negative binomial regression model is usually applied to estimate conditional expectations of the count outcome. The difference in conditional expectations under treatment and under control is then defined as the (conditional) treatment effect. While traditionally aggregates of these conditional treatment effects (e.g., average treatment effects) are computed by averaging over the empirical distribution, a recently proposed moment-based approach allows for computing aggregate effects as a function of distribution parameters. The moment-based approach makes it possible to control for (latent) multivariate normally distributed covariates and provides more reliable inferences under certain conditions. In this paper, we propose three different ways to account for non-normally distributed continuous covariates in this approach: (a) an alternative, known non-normal distribution, (b) a plausible factorization of the joint distribution, and (c) an approximation using finite Gaussian mixtures. A saturated model is used for categorical covariates, making a distributional assumption obsolete. We further extend the moment-based approach to allow for multiple treatment conditions and the computation of conditional effects for categorical covariates. An illustrative example highlighting the key features of our extension is provided.

4.1 Introduction

The evaluation of treatment effects on count outcomes is quite common in the social and health sciences. Often, covariates are included in the analysis using a negative binomial regression model (Garrett et al., 2018; Hittner et al., 2016; Jobe et al., 2001; Mazerolle et al., 2018; Nusser & Weinert, 2017; Schaumburg & Flynn, 2017; Sridharan, Shoda, Heffner, & Bricker, 2019), that is, the conditional expectation for the count outcome is logarithmically linked to the treatment variable and the covariates. The difference in conditional expectations under treatment vs. under control is then defined as the (conditional) treatment effect. Traditionally, aggregates of conditional treatment effects (e.g., the average treatment effect) are computed by averaging over the empirical (joint) distribution of the covariates (Greene, 2007).

Recently, a moment-based approach was proposed by Kiefer and Mayer (2019, *in press*) enabling aggregated effects to be computed using parameters of the (joint) distribution of the covariates, for example, means, variances, and covariances. In this approach, moment-generating functions are used as closed-form solutions for improper integrals over the covariates' distribution. The moment-based approach has two major advantages over the traditional empirical distribution approach. First, the covariate side of the effect calculation is stochastic instead of fixed, which enables more accurate statistical inferences about conditional and average treatment effects. Second, it allows common factor models for the covariates, meaning that not directly observed (latent) covariates can be accounted for as well. As the empirical distribution of latent variables is not observed, the traditional approach cannot deal with latent variables and would require fallible substitutes (i.e., sum scores, factor scores).

However, the extended moment-based approach as suggested by Kiefer and Mayer (*in press*) has an Achilles' heel: covariates are assumed to be multivariate normally distributed within the treatment groups. In practical settings, this assumption is often violated, as observed variables in real datasets in the social and health sciences frequently deviate from the normal distribution (Bono, Blanca, Arnau, & Gómez-Benito, 2017; Micceri, 1989). For example, Blanca, Arnau, Lopez-Montiel, Bono, and Bendayan (2013) examined 693 distributions from real psychological data and found that 74.4% presented slight or moderate deviations from the normal distribution, while 20% exhibited more extreme deviation. In a simulation study, Lei and Lomax (2005) investigated parameter and standard error bias in structural equation models when non-normal variables are introduced into normal distribution-based maximum likelihood estimation. They conclude that parameter estimation is sensitive to non-normal variables, but the bias introduced by slight non-normality is moderate. However, severe non-normality leads to substantial bias in loadings and structural parameters. Similarly, Kiefer and Mayer (2019) found in their examination of the univariate moment-based approach that misspecification of the covariates' distribution can introduce bias into the average effect estimation.

Thus, in this paper, we propose several extensions of the moment-based approach. These

extensions aim to make the approach more flexible with regard to its distributional assumptions as well as to provide applied researchers with a more nuanced effect analysis. First, to properly account for non-normal covariates, we suggest differentiating between categorical covariates (e.g., gender, ethnicity) and non-normal continuous covariates (e.g., pretest count variable). We then describe four different possibilities to account for non-normal joint distributions. Second, we define and compute average and conditional treatment effects of interest, for example, the conditional effect given a treatment condition or given a gender. These kinds of more fine-grained effects (as opposed to the average treatment effect) enable researchers to examine treatment efficacy under different conditions in a more nuanced way. Third, we consider the case in which multiple treatment conditions are to be examined, that is, treatment effects can be modeled and examined for more than one treatment condition (in comparison to a control group).

This paper is structured as follows: First, we provide non-parametric definitions of treatment effects on count outcomes based on multiple categorical and continuous covariates. Specifically, we will introduce the average treatment effect as well as several conditional treatment effects. Second, we derive how the moment-based approach can be utilized to compute these effects differentiating between a solution for categorical covariates and three possible approaches for non-normal continuous covariates. Third, we briefly present the negative binomial multigroup structural equation model as the statistical framework for simultaneously estimating the parameters of the negative binomial regression and the covariates' distribution. Finally, the moment-based approach is applied to a real data sample to highlight the key features of our proposed extensions.

4.2 Definition and Terminology of Treatment Effects

We start by introducing definitions of average and conditional treatment effects on a count outcome variable Y . Let the discrete treatment variable X have $p+1$ levels denoted with values $x = 0, 1, \dots, p$. We will take $X = 0$ as reference or control group and, thus, $t = 1, \dots, p$ are the values of the remaining treatment conditions. Furthermore, we will consider a single (unfolded) categorical variable K with $j+1$ levels and values $k = 0, 1, \dots, j$, and a vector of (latent) variables $\boldsymbol{\xi} = (1, \xi_1, \xi_2, \dots, \xi_q)$ with $z = 1, \dots, q$ denoting the $(z+1)$ th element of $\boldsymbol{\xi}$. For this set of variables, the following parameterization of $E(Y|X, K, \boldsymbol{\xi})$ always holds:

$$E(Y|X, K, \boldsymbol{\xi}) = g_0(K, \boldsymbol{\xi}) + \sum_{t=1}^p g_t(K, \boldsymbol{\xi}) \cdot I_{X=t} \quad (4.1)$$

$$= \sum_{x=0}^p \sum_{k=0}^j \exp[f_{xk}(\boldsymbol{\xi})] \cdot I_{X=x} \cdot I_{K=k} \quad (4.2)$$

In Equation (4.1), the conditional expectation of Y is decomposed into an intercept function g_0 and an effect function g_t for treatment condition t (cf. Mayer et al., 2016; Steyer

& Nagel, 2017). As Y is a count variable, the conditional expectation $E(Y|X, K, \boldsymbol{\xi})$ can always be presented as in Equation (4.2). Note that we use a regression with a logarithmic link function for each combination of X and K , that is, the function $f_{xk}(\boldsymbol{\xi})$ is group-specific:

$$E(Y|X = x, K = k, \boldsymbol{\xi}) = \exp[f_{xk}(\boldsymbol{\xi})]$$

We use group-specific functions $f_{xk}(\boldsymbol{\xi})$, because the effect computations later in this article will be based on a multigroup model.

4.2.1 Conditional Effects Function

The intercept function $g_0(K, \boldsymbol{\xi})$ reflects the expected values for the control group $X = 0$

$$\begin{aligned} g_0(K, \boldsymbol{\xi}) &= E(Y|X = 0, K, \boldsymbol{\xi}) \\ &= \sum_{k=0}^j \exp[f_{0k}(\boldsymbol{\xi})] \cdot I_{K=k} \end{aligned}$$

and the p effect functions $g_t(K, \boldsymbol{\xi})$ represent the expected increase/decrease in Y due to treatment t , that is, the difference between the conditional expectation of Y under treatment t compared to the control group:

$$\begin{aligned} g_t(K, \boldsymbol{\xi}) &= E(Y|X = t, K, \boldsymbol{\xi}) - E(Y|X = 0, K, \boldsymbol{\xi}) \\ &= \sum_{k=0}^j (\exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})]) \cdot I_{K=k} \end{aligned}$$

The effect function $g_t(K, \boldsymbol{\xi})$ gives a conditional treatment effect for any value of the covariates K and $\boldsymbol{\xi}$. The conditional expectations $E(Y|X = t, K, \boldsymbol{\xi})$ reflect the expected outcome given a treatment condition t . For example, $E(Y|X = 2, K = 1, \xi_1 = 1)$ reflects the expected count of Y for a individual with $K = 1$ who received training $X = 2$ and had a pretest score of $\xi_1 = 1$.

In practical settings, a parameterization is needed for $f_{xk}(\boldsymbol{\xi})$, which we will discuss at the end of this section. For a causal interpretation of the average and conditional effects, it is crucial that the effect functions $g_t(K, \boldsymbol{\xi})$ be causally unbiased. Steyer et al. (2014) provide an overview of causality conditions ensuring unbiasedness of the effect function, one of which is the independence of the treatment variable from all potential confounders (e.g., randomized treatment assignment as in the ACTIVE study).

4.2.2 Average Treatment Effect

The average treatment effect of treatment $X = t$ compared to control group $X = 0$ is defined as the unconditional expectation of the effect function

$$\begin{aligned} AE_{t0} &= E[g_t(K, \boldsymbol{\xi})] \\ &= E \left[\sum_{k=0}^j (\exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})]) \cdot I_{K=k} \right] \\ &= \sum_{k=0}^j \int_{\boldsymbol{\xi}} (\exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})]) \cdot p_{K,\boldsymbol{\xi}}(k, \boldsymbol{\xi}) d\boldsymbol{\xi} \end{aligned}$$

and therefore reflects the average over the conditional effects given all values of the covariates. The average effect AE_{t0} gives the expected effect for a random person assigned to treatment condition $X = t$ compared to the control group $X = 0$.

4.2.3 Conditional Effects given X and K

In addition to the average effect, more fine-grained aggregates of the effect function also exist. For example, the treatment effect for female participants can be examined. This is represented as a conditional effect given a value of the categorical covariate $K = k$ and averages over the $K = k$ -conditional distribution of the continuous covariates

$$\begin{aligned} CE_{t0;K=k} &= E[g_t(K, \boldsymbol{\xi}) | K = k] \\ &= \int_{\boldsymbol{\xi}} \exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})] \cdot p_{\boldsymbol{\xi}|K}(\boldsymbol{\xi} | k) d\boldsymbol{\xi} \end{aligned}$$

For example, if K represents a participant's gender, then $CE_{t0;K=k}$ provides the aggregated conditional effect of treatment $X = t$ for all persons with gender $K = k$.

In a similar vein, the conditional effect given a treatment condition is defined by

$$\begin{aligned} CE_{t0;X=x} &= E[g_t(K, \boldsymbol{\xi}) | X = x] \\ &= \sum_{k=0}^j \int_{\boldsymbol{\xi}} (\exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})]) \cdot p_{K,\boldsymbol{\xi}|X}(k, \boldsymbol{\xi} | x) d\boldsymbol{\xi} \end{aligned}$$

These effects are of interest when the X -conditional covariate distribution $p_{K,\boldsymbol{\xi}|X}$ differs among treatment groups, as is the case, for example, in observational studies (i.e., without randomization) or in so-called broken experiments ([Sagarin et al., 2014](#)). For an overview, see [Geneletti and Dawid \(2011\)](#). In properly randomized trials, however, we would expect the conditional effects given a treatment condition to be equivalent to the corresponding average treatment effect.

Finally, the conditional effects above can also be combined to give the conditional effect

given a value k of K and a treatment condition $X = x$:

$$\begin{aligned} CE_{t0;X=x,K=k} &= E[g_t(K, \boldsymbol{\xi}) | X = x, K = k] \\ &= \int_{\boldsymbol{\xi}} \exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})] \cdot p_{\boldsymbol{\xi}|X,K}(\boldsymbol{\xi}|x, k) d\boldsymbol{\xi} \end{aligned}$$

Again, for a randomized controlled trial, we would expect these effects to be equivalent to the corresponding treatment effects given a value of the categorical covariate.

4.2.4 Regression Parameterization and Covariates Factorization

At this point, we have presented non-parametrical definitions of average and conditional effects on count outcomes. In order to derive empirically estimable quantities, we need to introduce a parameterization for the $f_{xk}(\boldsymbol{\xi})$ functions and an applicable factorization of the mixed joint density of the categorical covariate K and the continuous covariates $\boldsymbol{\xi}$.

We choose a linear parameterization of the function $f_{xk}(\boldsymbol{\xi})$:

$$f_{xk}(\boldsymbol{\xi}) = \boldsymbol{\alpha}'_{xk} \boldsymbol{\xi} \quad (4.3)$$

where $\boldsymbol{\alpha}_{xk} = (\alpha_{xk0}, \alpha_{xk1}, \dots, \alpha_{xkq})'$ is a real-valued vector of regression coefficients with length $q + 1$. This corresponds to the parameterization widely used in count regression models, for example, Poisson or negative binomial regression models. Hence, a count regression is specified for each combination of XK

As can be seen in the previous subsection, the various average and conditional treatment effects require unconditional, but also X -, K -, and XK -conditional distributions of the covariates K and $\boldsymbol{\xi}$. Hence, we provide a factorization of the mixed joint distribution $p_{K,\boldsymbol{\xi}}(k, \boldsymbol{\xi})$ allowing us to easily derive all of the aforementioned unconditional and conditional distributions:

$$p_{K,\boldsymbol{\xi}}(k, \boldsymbol{\xi}) = \sum_{x=0}^p p_{X,K,\boldsymbol{\xi}}(x, k, \boldsymbol{\xi}) \quad (4.4)$$

$$= \sum_{x=0}^p P(X = x, K = k) \cdot p_{\boldsymbol{\xi}|X,K}(\boldsymbol{\xi}|x, k) \quad (4.5)$$

In Equation (4.4), the joint distribution of the covariates K and $\boldsymbol{\xi}$ is defined as the marginal distribution of the joint distribution of the treatment variable X , the categorical covariate K , and the continuous covariate $\boldsymbol{\xi}$. In Equation (4.5), the joint distribution is decomposed into a categorical group part $P(X = x, K = k)$ and a group-conditional density $p_{\boldsymbol{\xi}|X,K}(\boldsymbol{\xi}|x, k)$ of the continuous variables. The choice of a factorization is without loss of generality for our approach, because alternative factorization can always be transformed into our factorization. We provide more information on this aspect in the discussion and in Appendix 4.A. However, the factorization enables us to use a multigroup approach within the statistical framework.

Following the linear parameterization of the function $f_{xk}(\boldsymbol{\xi})$ and factorization of the joint distribution $p_{K,\boldsymbol{\xi}}(k, \boldsymbol{\xi})$, the average treatment effect, for example, can be computed as

$$\begin{aligned} AE_{t0} &= \sum_{k=0}^j \int_{\boldsymbol{\xi}} (\exp[f_{tk}(\boldsymbol{\xi})] - \exp[f_{0k}(\boldsymbol{\xi})]) \cdot p_{K,\boldsymbol{\xi}}(k, \boldsymbol{\xi}) d\boldsymbol{\xi} \\ &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \underbrace{\int_{\boldsymbol{\xi}} (\exp[\boldsymbol{\alpha}'_{tk}\boldsymbol{\xi}] - \exp[\boldsymbol{\alpha}'_{0k}\boldsymbol{\xi}]) \cdot p_{\boldsymbol{\xi}|K,X}(\boldsymbol{\xi}|x, k) d\boldsymbol{\xi}}_{\text{integration part of average effect}} \end{aligned} \quad (4.6)$$

Note that the parameterization and factorization are chosen with regard to the statistical framework for parameter estimation we will discuss later on, that is, negative binomial regression and multigroup structural equation models. As stated before, for a causal interpretation of the treatment effects, the effect function $g_t(K, \boldsymbol{\xi})$ must be causally unbiased. Thus, applied researchers should carefully evaluate whether the parameterization and factorization described here fits their hypotheses and assumptions.

4.3 Moment-Based Approach Assuming Multivariate Normality

In the previous section, we introduced a distinction between categorical and continuous covariates. Such a distinction has been previously proposed for treatment effect computation, for example, in the EffectLiteR approach by Mayer et al. (2016). A saturated model is used, that is, the probability of each occurring value of the categorical covariate is examined. Thus, no further distributional assumption for the categorical covariate is required.

However, including continuous covariates requires specification of the XK -conditional distribution $p_{\boldsymbol{\xi}|K,X}$ of $\boldsymbol{\xi}$. In their extension of the moment-based approach, Kiefer and Mayer (in press) suggest assuming X -conditional multivariate normality. Adapting this notion to our distinction between categorical covariates K and continuous covariates $\boldsymbol{\xi}$ and assuming XK -conditional normality, that is, $\boldsymbol{\xi}|X = x, K = k \sim \mathcal{N}(\boldsymbol{\mu}_{xk}, \boldsymbol{\Sigma}_{xk})$, the integration part of the average treatment effect in Equation (4.6) can be substituted with moment-generating functions:

$$\begin{aligned} &\int_{\boldsymbol{\xi}} (\exp[\boldsymbol{\alpha}'_{tk}\boldsymbol{\xi}] - \exp[\boldsymbol{\alpha}'_{0k}\boldsymbol{\xi}]) \cdot p_{\boldsymbol{\xi}|K,X}(\boldsymbol{\xi}|x, k) d\boldsymbol{\xi} \\ &= \exp \left[\boldsymbol{\alpha}'_{tk}\boldsymbol{\mu}_{xk} + \frac{\boldsymbol{\alpha}'_{tk}\boldsymbol{\Sigma}_{xk}\boldsymbol{\alpha}'_{tk}}{2} \right] - \exp \left[\boldsymbol{\alpha}'_{0k}\boldsymbol{\mu}_{xk} + \frac{\boldsymbol{\alpha}'_{0k}\boldsymbol{\Sigma}_{xk}\boldsymbol{\alpha}'_{0k}}{2} \right] \end{aligned}$$

Sometimes, it might suffice to account for the case of categorical covariates in addition to normally distributed continuous covariates to obtain unbiased parameter and effect estimates.

4.4 Accounting for Non-Normal Continuous Covariates

In the following subsections, we present three different ways to incorporate non-normal continuous covariates into the moment-based approach. All three ways follow the basic idea of the moment-based approach, that is, (at least partly) substituting improper integrals with moment-generating functions. The suggested solutions depend on the level of information available for the non-normal continuous variables: (1) an alternative, non-normal joint distribution is known, (2) a plausible factorization of the joint distribution can be constructed, and (3) an approximation of the covariates' joint distribution is required. Note that we do not suggest that one of these approaches is generally preferable to the others. Rather, which approach to use depends on the concrete data situation one is confronted with.

4.4.1 Case 1: Known Non-Normal Joint Distribution

The first and probably simplest case is a known alternative non-normal joint distribution for $\xi|X, K$. In this case, we can substitute the density and moment-generating functions for effect computation with the ones from the alternative distribution. For example, if we want to account for a slight skew in our continuous variables, the multivariate skew-normal distribution ([Azzalini & Valle, 1996](#)) is a viable alternative to the normal distribution. Let the continuous covariates be skew-normally distributed with $\xi|X = x, K = k \sim \text{SN}(\zeta_{xk}, \Omega_{xk}, \Delta_{xk})$, then the integration part of the average effect in Equation (4.6) can be denoted with the respective moment-generating function given by [Gupta, Gonzalez-Farias, and Dominguez-Molina \(2004\)](#):

$$\begin{aligned} & \int_{\xi} (\exp[\alpha'_{tk}\xi] - \exp[\alpha'_{0k}\xi]) \cdot p_{\xi|K,X}(\xi|x, k) d\xi \\ &= 2 \cdot \left[\exp\left(\alpha'_{tk}\zeta_{xk} + \frac{\alpha'_{tk}\Omega_{xk}\alpha_{tk}}{2}\right) \frac{\Phi(\Delta_{xk}\Omega_{xk}\alpha_{tk}; I + \Delta\Omega\Delta'_{xk})}{\Phi(0; I + \Delta\Omega\Delta'_{xk})} \right. \\ & \quad \left. - \exp\left(\alpha'_{0k}\zeta_{xk} + \frac{\alpha'_{0k}\Omega_{xk}\alpha_{0k}}{2}\right) \frac{\Phi(\Delta_{xk}\Omega_{xk}\alpha_{0k}; I + \Delta\Omega\Delta'_{xk})}{\Phi(0; I + \Delta\Omega\Delta'_{xk})} \right] \end{aligned}$$

where $\Phi(\mu, \Sigma)$ denotes the cumulative distribution function of a multivariate normal distribution with mean vector μ and covariance matrix Σ . The parameters of the skew-normal distribution denote a location parameter ζ_{xk} , a scale matrix Ω_{xk} , and a slant matrix Δ_{xk} . For more information on the skew-normal distribution, see [Azzalini and Capitanio \(2009\)](#).

Another important scenario for known alternative distributions is conditional independence of the covariates $\xi|X, K$, because then the product of univariate distributions corresponds to the marginal distribution. For example, if we consider two continuous covariates ξ_1 and ξ_2 with XK -conditional independence $\xi_1 \perp\!\!\!\perp \xi_2|X, K$, the corresponding density can be decomposed $p_{\xi_1, \xi_2} = p_{\xi_1} \cdot p_{\xi_2}$ and thus, the moment-generating function can be denoted

as product of univariate moment-generating functions:

$$M_{\xi_1, \xi_2 | X, K}(t_1, t_2) = M_{\xi_1}(t_1) \cdot M_{\xi_2}(t_2)$$

where t_1, t_2 are the evaluation points of the moment-generating functions. See [Kiefer and Mayer \(2019\)](#) for an overview of univariate moment-generating functions and their performance within the moment-based approach.

4.4.2 Case 2: Factorization of Joint Distribution

Sometimes, it might be impossible to find a suitable joint distribution, and the assumption of conditional independence might be too strong. A notable example of this case was chosen for our illustrative example: a continuous latent variable (i.e., depression) and a discrete, yet non-categorical count variable (i.e., baseline count of correctly answered items). To our knowledge, no joint distributions for count and continuous variables have been proposed to date. However, when examining count outcome variables, researchers often wish to account for the respective baseline count as well.

We suggest a practical workaround for these cases: finding a plausible factorization of the joint distribution. For example, this might be decomposing the joint distribution of ξ into a marginal and a conditional distribution, that is,

$$p_{\xi | X, K} = p_{\xi_m | X, K} \cdot p_{\xi_c | X, K, \xi_m}$$

where we specify the marginal distribution of $\xi_m \subset \xi$ and the conditional distribution of $\xi_c \subset \xi$ given ξ_m . In general, this factorization makes it possible to simplify the integration part of the average effect in Equation (4.6) as

$$\begin{aligned} & \int_{\xi} (\exp[\alpha'_{tk} \xi] - \exp[\alpha'_{0k} \xi]) \cdot p_{\xi | K, X}(\xi | x, k) d\xi \\ &= \int_{\xi_m} \left(\exp[\alpha'_{tk; m} \xi_m] M_{\xi_c | \xi_m}(\alpha_{tk; c}) - \exp[\alpha'_{0k; m} \xi_m] M_{\xi_c | \xi_m}(\alpha_{0k; c}) \right) \cdot p_{\xi_m | X, K}(\xi_m) \cdot d\xi_m \end{aligned}$$

The factorization approach is challenging to integrate into the moment-based approach, because it does not yield a comprehensive moment-generating function for the factorized joint distribution. Consequently, a combination of moment-generating functions and numerical integration is required for effect estimation. In our illustrative example, we will further delineate these computations for the factorization approach.

In practical terms, the construction of a conditional distribution for some covariates can be achieved using a regression approach, that is, a parameterization of $E(\xi_c | X = x, K = k, \xi_m)$. In our illustrative example, this relation will be estimated using a regression with a logarithmic link function. From a causal perspective, such a regression can be problematic, because baseline variables do not necessarily have a temporal order, and we do not suggest that one baseline variable *causes* another baseline variable. Thus, the

factorization approach is a technical workaround and its regression parameters are not necessarily of interest.

4.4.3 Case 3: Approximation with Finite Gaussian Mixtures

Finally, when no alternative joint distribution or plausible factorization can be found, it is possible to approximate the joint distribution of ξ . One possible approach is to approximate the non-normal joint distribution of ξ by using a finite mixture of M multivariate normal distributions. That is, $p_{\xi|K,X}(\xi|x,k) = \sum_{m=1}^M w_m \cdot p_{\xi|K,X;m}(\xi|x,k;m)$ where w_m are weights and $\xi|X=x, K=k; m \sim \mathcal{N}(\mu_{xk;m}, \Sigma_{xk;m})$. In practice, this approximation of $\xi|X=x, K=k$ can be estimated using a finite mixture or latent class approach (McLachlan & Peel, 2000). Researchers can control the degree of approximation by specifying the number M of latent classes, where only the expected values $\mu_{xk;m}$ and the variance $\Sigma_{xk;m}$ are allowed to vary among latent classes. This approach has previously been applied in the computation of average and conditional effects in a non-linear regression setting by Mayer et al. (2017) and can be estimated using nonlinear structural equation mixture models (Kelava & Brandt, 2014; Kelava et al., 2014). However, it is notable that an increasing number of latent classes M can lead to convergence problems in maximum likelihood estimation.

With respect to effect estimation, the finite mixture approach yields a comprehensive moment-generating function for $\xi|X, K$, as we can use a weighted sum of normal moment-generating functions for the integration part from Equation (4.6):

$$\begin{aligned} & \int_{\xi} (\exp[\alpha'_{tK}\xi] - \exp[\alpha'_{0K}\xi]) \cdot p_{\xi|K,X}(\xi|x,k) d\xi \\ &= \sum_{m=1}^M w_m \exp \left[\alpha'_{tk} \mu_{tk;m} + \frac{\alpha'_{tk} \Sigma_{tk;m} \alpha'_{tk}}{2} \right] - \exp \left[\alpha'_{0k} \mu_{0k;m} + \frac{\alpha'_{0k} \Sigma_{0k;m} \alpha'_{0k}}{2} \right] \end{aligned}$$

which is similar to the original extended moment-based approach by Kiefer and Mayer (in press), with the exception of the summation over the latent classes.

4.5 Negative Binomial Multigroup Structural Equation Model

In this section, we introduce the negative binomial multigroup structural equation model as a statistical framework for parameter and effect estimation. The NB-MG-SEM offers a maximum likelihood estimation of the parameters involved in effect estimation. The model involves the count outcome variable Y , the categorical treatment variable X with p levels, the (unfolded) categorical covariate K with j levels, the vector of continuous covariates $\xi = (1, \xi_1, \dots, \xi_q)$ containing latent covariates measured by observed variables $z = (Z_1, Z_2, \dots, Z_s)$. The negative binomial multigroup structural equation model consists

of (at least) the following parts:

$$\mathbf{z} = \boldsymbol{\nu} + \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon} \quad (4.7)$$

$$\mu_Y = \exp(\boldsymbol{\alpha}'_{xk}\boldsymbol{\xi}) \quad (4.8)$$

$$\log(n_{xk}) = \kappa_{xk} \quad (4.9)$$

where $\boldsymbol{\nu}$ is a vector of measurement intercepts; $\mathbf{\Lambda}$ is a matrix of loadings; $\boldsymbol{\varepsilon}$ is a vector of measurement error variables with mean zero and covariance matrix $\boldsymbol{\Theta}_{xk}$; μ_Y is the conditional expectation of the count outcome; $\boldsymbol{\alpha}_{xk}$ is a vector of regression coefficients; and κ_{xk} is a parameter for the log-transformed expected group frequency of group xk . The probability for a group xk can be computed with $P(X = x, K = k) = \frac{\exp(\kappa_{xk})}{\sum_{x^*=0}^p \sum_{k^*=0}^j \exp(\kappa_{x^*k^*})}$.

The NB-MG-SEM presented in Equations (4.7) to (4.9) is a least common denominator with regard to the cases of non-normally distributed covariates identified in the previous section and can be extended with respect to the given case at hand. In our illustrative example, we will add another regression with a logarithmic link function specifying the relation among two covariates. We provide detailed information on the maximum likelihood estimation of the NB-MG-SEM for our illustrative example in Appendix 4.B. For brevity, we present the four main parts and assumptions of the likelihood functions constituting the joint distribution of the considered variables:

$$\begin{aligned} n_{xk} &\sim \mathcal{P}(\exp(\kappa_{xk})) \\ Y|\boldsymbol{\xi}, X = x, K = k &\sim \text{NB}(\mu_Y, \phi_{xk}) \\ \mathbf{z}|\boldsymbol{\xi}, X = x, K = k &\sim \mathcal{N}(\boldsymbol{\nu} + \mathbf{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Theta}_{xk}) \\ \boldsymbol{\xi}|X = x, K = k & \end{aligned}$$

The joint distribution of the covariates $\boldsymbol{\xi}|X = x, K = k$ depends on which of the three aforementioned cases is applied. In our illustrative example, we will factorize $\boldsymbol{\xi}|X, K$ into a marginal and a conditional distribution (i.e., Case 2 scenario), where the conditional distribution depends on a μ_{ξ_2} and the parameters of the marginal distribution are implied by the measurement model.

4.6 Illustrative Example

To illustrate the use of our extension to the moment-based approach, we use data from the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) study (Ball et al., 2002; Jobe et al., 2001). The ACTIVE study is a large randomized controlled trial designed to examine the effectiveness of cognitive interventions among older adults. We will investigate the effects of these interventions while controlling for the (non-normal) count pretest score, (latent) baseline depression, and participants' gender. As count outcome Y , we investigate post-test performance on an inductive reasoning assessment (i.e., letter sets). Letter sets evaluate how well an individual can recognize a pattern among several

sets of letters. Thus, Y reflects the count of correctly answered items.

In this article, we do not present a comprehensive analysis of the ACTIVE study. Instead, the primary goal of this paper is to illustrate how non-normally distributed covariates can be integrated into the moment-based approach when estimating average and conditional treatment effects. The NB-MG-SEM and treatment effects were estimated using R (R Core Team, 2018) and the lavaan package (Rosseel, 2012). An annotated version of our code is provided in the supplementary materials¹. We used listwise deletion in our analyses in order to keep the supplementary R code accessible for interested readers. However, we also ran a full information maximum likelihood estimation for our model (using Mplus; Muthén & Muthén, 1998-2015) which yielded similar results.

4.6.1 Sample

The total sample size for our analysis was $N = 1754$. The participants were randomly assigned to one of four conditions: a no-contact control group ($X = 0$, $N = 424$), a memory training ($X = 1$, $N = 443$), a reasoning training ($X = 2$, $N = 438$), and a speed of processing training $X = 3$, $N = 449$. Each treatment condition consisted of a ten-session training intervention. In a baseline assessment, the participants took several cognitive functioning tests and completed a self-report questionnaire of psychological measures. The post-test assessment was conducted in the first ten days after the last training session. The participants were predominantly female (76.6%, $K = 1$).

4.6.2 Measures

Depression Depressive symptoms were assessed using the 12-item version of the Center for Epidemiological Studies-Depression Scale (CESD-12; Radloff, 1977). Participants rated the frequency of several depressive symptoms (e.g., feeling sad) during the last week on a 4-point scale from 0 = *never* to 3 = *5-7 days*. We modeled baseline depression as latent variable ξ_1 , measured by three parcels (i.e., sum scores of four items each; Z_{11} , Z_{12} , Z_{13}). We used random item parcels for simplicity in our illustrative example. For cautions about the use of parcels, see Marsh et al. (2013) and Sterba and Rights (2016).

Letter Sets The count of correctly answered items on a letter sets task was used as outcome Y (i.e., post-test score) and covariate ξ_2 (i.e., pretest score) in our analysis. Letter sets evaluate how well an individual can find rules or patterns that make different sets of letters alike in some way. Each problem had five sets of letters with four letters in each set. Participants were given 15 different problems and had seven minutes to complete the task.

¹We also provide a R package which is able to carry out the computations in this illustrative example. However, to ensure a blind review, we chose not to provide a name or link, but only the annotated R code for now.

4.6.3 Model

Our effect analysis was based on a multigroup structural equation model with a group-invariant linear measurement model and a structural model with two linear predictors, logarithmic link functions, and a negative binomial distribution with overdispersion parameter ϕ fixed to zero (i.e., a Poisson distribution) for the respective dependent variable². The measurement model was chosen to be τ -congeneric, as we had no a priori assumption about the true scores (e.g., τ -equivalence). The first indicator was chosen as reference indicator. The measurement model is expressed as:

$$\begin{pmatrix} Z_{11} \\ Z_{12} \\ Z_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ \nu_{12} \\ \nu_{13} \end{pmatrix} + \begin{pmatrix} 1 \\ \lambda_{12} \\ \lambda_{13} \end{pmatrix} \cdot \xi_1 + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \end{pmatrix}$$

The latent variable ξ_1 was assumed to have an XK -conditional normal distribution $\xi_1|X = x, K = k \sim \mathcal{N}(\mu_{xk}, \sigma_{xk})$.

Our structural model with two linear predictors, logarithmic link functions and Poisson distribution for the respective dependent variable was:

$$\begin{aligned} E(Y|X = x, K = k, \xi_1, \xi_2) &= \exp(\alpha_{xk0} + \alpha_{xk1}\xi_1 + \alpha_{xk2}\xi_2) \\ E(\xi_2|X = x, K = k, \xi_1) &= \exp(\gamma_{xk0} + \gamma_{xk1}\xi_1) \end{aligned}$$

Note that the second regression $E(\xi_2|X = x, K = k, \xi_1)$ is specified for technical reasons, that is, to model a factorized joint distribution of ξ_1 and ξ_2 . We do not assume a causal relation, where the values of ξ_2 are predetermined by ξ_1 . Hence, it would also be possible to factorize the joint distribution vice versa, with $E(\xi_1|X = x, K = k, \xi_2)$. We chose the first factorization, because it yields some facilitative properties for the numerical integration procedures required in effect estimation (i.e., Gauss-Hermite quadrature; for more information, see Appendix 4.C).

Finally, the model for group sizes n_{xk} was expressed as

$$\kappa_{xk} = \log(n_{xk})$$

The whole model is displayed in Figure 4.1. Note that the regression with a logarithmic link function are indicated by curved arrows in contrast to straight lines for linear regressions.

4.6.4 Average and Conditional Treatment Effects

With regard to our three proposed scenarios of non-normal continuous covariates, the specification of a baseline count variable and a continuous variable falls under Case 2: as no

²In our analysis, we fixed the overdispersion parameter $\phi = 0$ for technical reasons, as the log-likelihood estimation did not converge for non-zero overdispersion parameters. We will discuss the issue of estimation difficulties at the end of the paper.

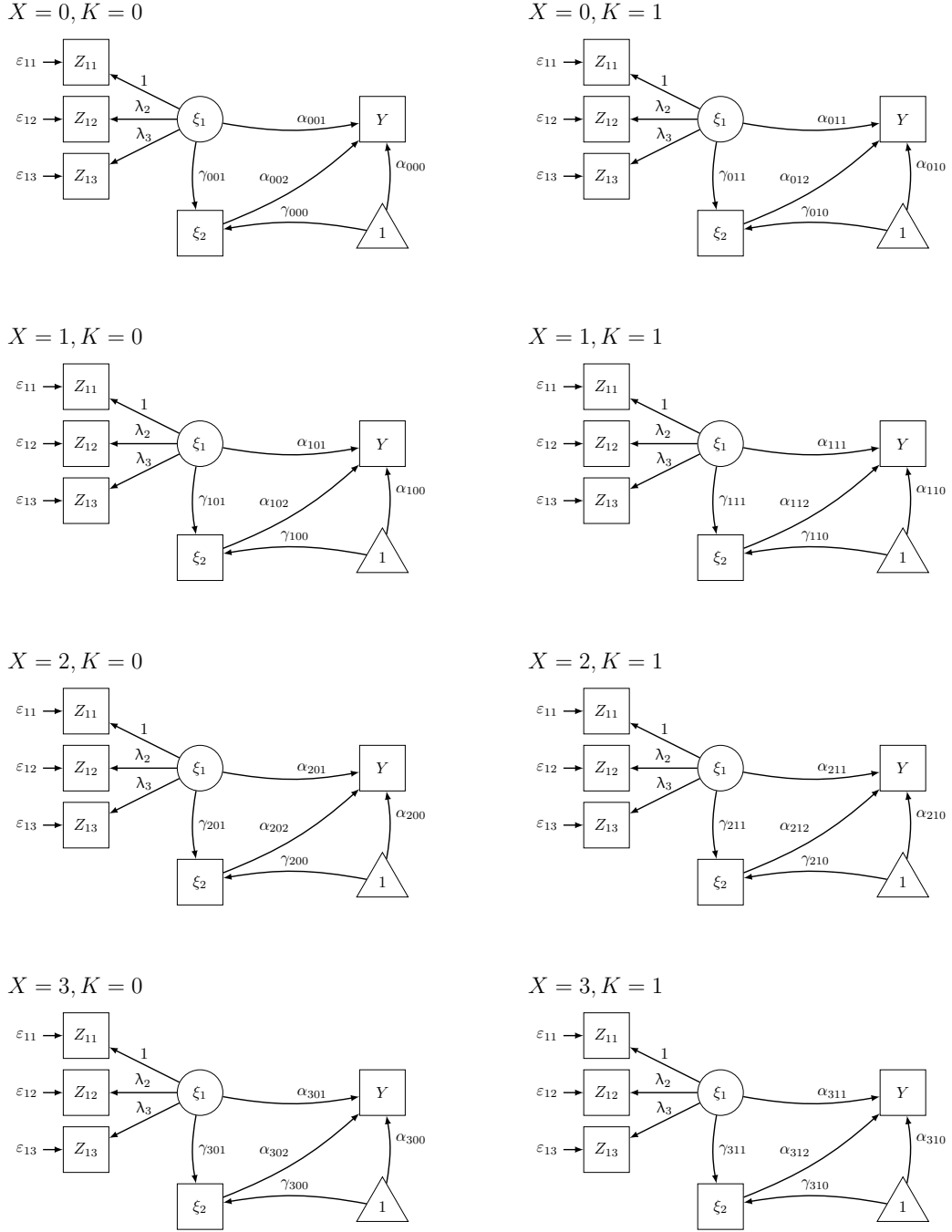


Figure 4.1

Path diagram for our illustrative example depicting the group invariant measurement models for latent baseline depression ξ_1 and the baseline count of correctly answered items ξ_2 , the structural model specifying the regressions with a logarithmic link function (indicated by curved arrows) of the post-test count of correctly answered items Y on the baseline variables ξ_1 and ξ_2 in each of the eight groups.

suitable joint distribution is available, a factorization of the joint distribution into a known marginal and a known conditional distribution can be specified. In the previous section, we stated that we assumed latent baseline depression is XK -conditionally marginally normally distributed $\xi_1|X = x, K = k \sim \mathcal{N}(\mu_{xk}, \sigma_{xk})$ and is related to the baseline count of correctly answered items via a Poisson regression with $\xi_2|X = x, K = k, \xi_1 \sim \mathcal{P}(\lambda_{xk})$ with $\lambda_{xk} = \exp(\gamma_{xk0} + \gamma_{xk1}\xi_1)$. In this scenario, the average treatment effect can be computed as

$$ATE = \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \sum_{h=1}^H w_h [\exp(\alpha_{tk0} + \alpha_{tk1}\xi_{1h}^*) \cdot M_{\xi_2|\xi_1, K, X}(\alpha_{tk2}) \\ - \exp(\alpha_{0k0} + \alpha_{0k1}\xi_{1h}^*) \cdot M_{\xi_2|\xi_1, K, X}(\alpha_{0k2})]$$

where $\xi_{1h}^* = \sqrt{2\sigma_{xk}^2}a_h + \mu_{xk}$ and a_h, w_h are Gauss-Hermite quadrature points and weights. A detailed derivation of this formula is provided in Appendix 4.C. The moment-generating function for the Poisson distribution is:

$$M_{\xi_2}(\alpha_{xk2}) = \exp(\lambda_{xk}[\exp(\alpha_{xk2}) - 1]) \\ = \exp(\exp[\gamma_{xk0} + \gamma_{xk1}\xi_1][\exp(\alpha_{xk2}) - 1])$$

The conditional effects are then computed analogously to our definitions above. For example, the conditional effect given a value of the gender covariate K is computed:

$$CE_{t0;K=k} = \sum_{x=0}^p P(X = x, K = k) \cdot \sum_{h=1}^H w_h [\exp(\alpha_{tk0} + \alpha_{tk1}\xi_{1h}^*) M_{\xi_2|\xi_1, K, X}(\alpha_{tk2}) \\ - \exp(\alpha_{0k0} + \alpha_{0k1}\xi_{1h}^*) M_{\xi_2|\xi_1, K, X}(\alpha_{0k2})]$$

where $\xi_{1h}^* = \sqrt{2\sigma_{xk}^2}a_h + \mu_{xk}$ are integration points at which the function is evaluated and a_h, w_h are Gauss-Hermite quadrature points and weights.

4.6.5 Results

Currently, it is not possible to evaluate the complete model fit using a χ^2 -test or other fit statistics, as these are based on the χ^2 -value of the model implied covariance matrix for the observed variables. For models with a logarithmic link function, the implied covariance matrix is not an appropriate description of the dependencies among the variables due to the non-linearity.

Model parameters

The maximum likelihood estimates for the model parameters are given in Table 4.1. We will not discuss them in detail here, but rather present some notable findings to illustrate the interpretation of these parameters.

Pretest depression had no significant effect on the post-test count of correctly answered

Table 4.1: Maximum likelihood results for group-specific model parameters

Parameter	Estimate	SE	p	Parameter	Estimate	SE	p
$X = 0, K = 0$				$X = 0, K = 1$			
Estimates for $E(Y X = 0, K = 0, \xi_1, \xi_2)$				Estimates for $E(Y X = 0, K = 1, \xi_1, \xi_2)$			
α_{000}	1.299	0.114	< .001	α_{010}	1.221	0.071	< .001
α_{001}	-0.016	0.035	.648	α_{011}	-0.033	0.019	.083
α_{002}	0.102	0.013	< .001	α_{012}	0.103	0.008	< .001
Estimates for $E(\xi_2 X = 0, K = 0, \xi_1)$				Estimates for $E(\xi_2 X = 0, K = 1, \xi_1)$			
γ_{000}	1.929	0.066	< .001	γ_{010}	1.956	0.038	< .001
γ_{001}	-0.062	0.042	.140	γ_{011}	-0.110	0.020	< .001
Estimates for $\xi_1 X = 0, K = 0$				Estimates for $\xi_1 X = 0, K = 1$			
μ_{00}	1.375	0.121	< .001	μ_{01}	1.645	0.093	< .001
σ_{00}^2	1.217	0.207	< .001	σ_{01}^2	2.060	0.212	< .001
$X = 1, K = 0$				$X = 1, K = 1$			
Estimates for $E(Y X = 1, K = 0, \xi_1, \xi_2)$				Estimates for $E(Y X = 1, K = 1, \xi_1, \xi_2)$			
α_{100}	1.314	0.141	< .001	α_{110}	1.263	0.061	< .001
α_{101}	-0.014	0.041	.733	α_{111}	-0.013	0.016	.417
α_{102}	0.096	0.016	< .001	α_{112}	0.101	0.007	< .001
Estimates for $E(\xi_2 X = 1, K = 0, \xi_1)$				Estimates for $E(\xi_2 X = 1, K = 1, \xi_1)$			
γ_{100}	1.972	0.067	< .001	γ_{110}	1.876	0.035	< .001
γ_{101}	-0.050	0.042	.234	γ_{111}	-0.037	0.019	.052
Estimates for $\xi_1 X = 1, K = 0$				Estimates for $\xi_1 X = 1, K = 1$			
μ_{10}	1.365	0.122	< .001	μ_{11}	1.513	0.083	< .001
σ_{10}^2	1.089	0.202	< .001	σ_{11}^2	1.854	0.190	< .001
$X = 2, K = 0$				$X = 2, K = 1$			
Estimates for $E(Y X = 2, K = 0, \xi_1, \xi_2)$				Estimates for $E(Y X = 2, K = 1, \xi_1, \xi_2)$			
α_{200}	1.437	0.142	< .001	α_{210}	1.400	0.058	< .001
α_{201}	-0.051	0.032	.111	α_{211}	-0.022	0.015	.143
α_{202}	0.098	0.015	< .001	α_{212}	0.095	0.07	< .001
Estimates for $E(\xi_2 X = 2, K = 0, \xi_1)$				Estimates for $E(\xi_2 X = 2, K = 1, \xi_1)$			
γ_{200}	2.106	0.061	< .001	γ_{210}	1.780	0.038	< .001
γ_{201}	-0.132	0.033	< .001	γ_{211}	0.004	0.018	.824
Estimates for $\xi_1 X = 2, K = 0$				Estimates for $\xi_1 X = 2, K = 1$			
μ_{20}	1.646	0.168	< .001	μ_{21}	1.672	0.085	< .001
σ_{20}^2	2.185	0.403	< .001	σ_{21}^2	1.871	0.191	< .001
$X = 3, K = 0$				$X = 3, K = 1$			
Estimates for $E(Y X = 3, K = 0, \xi_1, \xi_2)$				Estimates for $E(Y X = 3, K = 1, \xi_1, \xi_2)$			
α_{300}	1.294	0.116	< .001	α_{310}	1.265	0.064	< .001
α_{301}	0.013	0.030	.665	α_{311}	-0.010	0.018	.579
α_{302}	0.095	0.013	< .001	α_{312}	0.102	0.008	< .001
Estimates for $E(\xi_2 X = 3, K = 0, \xi_1)$				Estimates for $E(\xi_2 X = 3, K = 1, \xi_1)$			
γ_{300}	1.962	0.063	< .001	γ_{310}	1.844	0.038	< .001
γ_{301}	-0.085	0.037	.022	γ_{311}	-0.038	0.020	.058
Estimates for $\xi_1 X = 3, K = 0$				Estimates for $\xi_1 X = 3, K = 1$			
μ_{30}	1.465	0.149	< .001	μ_{31}	1.579	0.078	< .001
σ_{30}^2	1.859	0.353	< .001	σ_{31}^2	1.574	0.159	< .001

items in either group (e.g., $\alpha_{001} = -0.016$, 95% $CI = [-0.085, 0.053]$). Furthermore, pretest depression was significantly negatively related to the pretest count of correctly answered items in three groups (i.e., $\gamma_{011} = -0.110$, 95% $CI = [-0.149, -0.071]$, $\gamma_{201} = -0.132$, 95% $CI = [-0.197, -0.067]$, and $\gamma_{301} = -0.085$, 95% $CI = [-0.158, -0.012]$). For example, for male participants in the reasoning training, each 1-unit change in pretest depression was linked to a (i.e., γ_{201} , $\exp(-0.132) = 0.88$) 12% decrease in correctly answered items at baseline. In the other five groups, this relationship was not significant.

Pretest count of correctly answered items was a significant positive predictor of post-test count of correctly answered items in all groups. For example, for female participants in the memory group, each additional correctly answered item at baseline was linked with a 11% increase in correctly answered items at posttest (i.e., α_{012} , $\exp(0.103) = 1.11$). The intercept coefficients reflect the expected posttest count of correctly answered items for a male person with depression score ($\xi_1 = 0$) and zero correctly answered items at baseline ($\xi_2 = 0$). For example, in the reasoning training, (i.e., $\alpha_{200} = 1.437$), the expected count is $\hat{E}(Y|X = 2, K = 0, \xi_1 = 0, \xi_2 = 0) = \exp(1.437) = 4.21$.

Average effects

An overview of all average and conditional treatment effects estimated for our illustrative example is given in Table 4.2. Remember that the letter sets test was part of a cognitive assessment measuring reasoning performance. Hence, we would expect the reasoning training in particular to have a significant effect, while the memory and speed of processing trainings might not necessarily affect reasoning performance. In line with this, the average treatment effect of the memory training ($\widehat{AE}_{10} = 0.243$, 95% $CI = [-0.095, 0.585]$, $ES = 0.082$) and the average treatment effect of the speed of processing training ($\widehat{AE}_{30} = 0.335$, 95% $CI = [-0.008, 0.678]$, $ES = 0.115$) on the count of correctly answered items were not significant. The average treatment effect of the reasoning training, however, was significant ($\widehat{AE}_{20} = 0.849$, 95% $CI = [0.497, 1.201]$, $ES = 0.291$), that is, participants in this condition correctly answered 0.849 items more on average compared to baseline.

Conditional effects given K

The conditional treatment effects given a gender are given in detail in Table 4.2. Here, we examine differential treatment effects depending on the participants' gender.

For the memory training, we found a slightly negative treatment effect for male participants ($\widehat{CE}_{10;K=0} = -0.151$, 95% $CI = [-0.878, 0.575]$, $ES = -0.052$) and a positive treatment effect for female participants $K = 1$ ($\widehat{CE}_{10;K=1} = 0.364$, 95% $CI = [-0.022, 0.751]$, $ES = 0.125$). While the effect sizes differed in direction and magnitude, both effects were not statistically significant.

The reasoning training had significant effects for both men and women. The conditional effect for female participants ($\widehat{CE}_{20;K=1} = 0.956$, 95% $CI = [0.557, 1.355]$, $ES = 0.327$)

Table 4.2: Estimated average and conditional effects of memory training ($X = 1$ vs. $X = 0$), reasoning training ($X = 2$ vs. $X = 0$), and speed of processing training ($X = 3$ vs. $X = 0$) compared to the control group

Effect	Estimate	SE	ES	Effect	Estimate	SE	ES
Estimated average and conditional effects of memory training							
Average Effect $\widehat{AE}_{10}$				Conditional effects $\widehat{CE}_{10;X=x,K=k}$			
$\widehat{AE}_{10}$	0.243	0.174	0.083	$\widehat{CE}_{10;X=0,K=0}$	-0.138	0.363	-0.047
Conditional effects $\widehat{CE}_{10;X=x}$				$\widehat{CE}_{10;X=0,K=1}$	0.357	0.197	0.122
$\widehat{CE}_{10;X=0}$	0.227	0.174	0.078	$\widehat{CE}_{10;X=1,K=0}$	-0.160	0.378	-0.055
$\widehat{CE}_{10;X=1}$	0.241	0.177	0.083	$\widehat{CE}_{10;X=1,K=1}$	0.358	0.200	0.123
$\widehat{CE}_{10;X=2}$	0.254	0.177	0.087	$\widehat{CE}_{10;X=2,K=0}$	-0.170	0.390	-0.058
$\widehat{CE}_{10;X=3}$	0.251	0.173	0.086	$\widehat{CE}_{10;X=2,K=1}$	0.379	0.197	0.130
Conditional effects $\widehat{CE}_{10;K=k}$				$\widehat{CE}_{10;X=3,K=0}$	-0.139	0.363	-0.048
$\widehat{CE}_{10;K=0}$	-0.151	0.370	-0.052	$\widehat{CE}_{10;X=3,K=1}$	0.363	0.196	0.124
$\widehat{CE}_{10;K=1}$	0.364	0.197	0.125				
Estimated average and conditional effects of reasoning training							
Average Effect $\widehat{AE}_{20}$				Conditional effects $\widehat{CE}_{20;X=x,K=k}$			
$\widehat{AE}_{20}$	0.849	0.180	0.291	$\widehat{CE}_{20;X=0,K=0}$	0.501	0.380	0.172
Conditional effects $\widehat{CE}_{20;X=x}$				$\widehat{CE}_{20;X=0,K=1}$	0.950	0.204	0.325
$\widehat{CE}_{20;X=0}$	0.832	0.181	0.285	$\widehat{CE}_{20;X=1,K=0}$	0.509	0.389	0.174
$\widehat{CE}_{20;X=1}$	0.858	0.184	0.294	$\widehat{CE}_{20;X=1,K=1}$	0.959	0.208	0.328
$\widehat{CE}_{20;X=2}$	0.855	0.182	0.293	$\widehat{CE}_{20;X=2,K=0}$	0.487	0.397	0.167
$\widehat{CE}_{20;X=3}$	0.852	0.179	0.292	$\widehat{CE}_{20;X=2,K=1}$	0.961	0.202	0.329
Conditional effects $\widehat{CE}_{20;K=k}$				$\widehat{CE}_{20;X=3,K=0}$	0.497	0.379	0.170
$\widehat{CE}_{20;K=0}$	0.499	0.381	0.171	$\widehat{CE}_{20;X=3,K=1}$	0.954	0.202	0.327
$\widehat{CE}_{20;K=1}$	0.956	0.203	0.327				
Estimated average and conditional effects of speed of processing training							
Average Effect $\widehat{AE}_{30}$				Conditional effects $\widehat{CE}_{30;X=x,K=k}$			
$\widehat{AE}_{30}$	0.335	0.175	0.115	$\widehat{CE}_{30;X=0,K=0}$	-0.090	0.366	-0.031
Conditional effects $\widehat{CE}_{30;X=x}$				$\widehat{CE}_{30;X=0,K=1}$	0.459	0.198	0.157
$\widehat{CE}_{30;X=0}$	0.316	0.175	0.108	$\widehat{CE}_{30;X=1,K=0}$	-0.113	0.382	-0.039
$\widehat{CE}_{30;X=1}$	0.334	0.179	0.114	$\widehat{CE}_{30;X=1,K=1}$	0.464	0.202	0.159
$\widehat{CE}_{30;X=2}$	0.350	0.182	0.120	$\widehat{CE}_{30;X=2,K=0}$	-0.101	0.395	-0.035
$\widehat{CE}_{30;X=3}$	0.342	0.174	0.117	$\widehat{CE}_{30;X=2,K=1}$	0.482	0.197	0.165
Conditional effects $\widehat{CE}_{30;K=k}$				$\widehat{CE}_{30;X=3,K=0}$	-0.084	0.367	-0.029
$\widehat{CE}_{30;K=0}$	-0.097	0.374	-0.033	$\widehat{CE}_{30;X=3,K=1}$	0.464	0.197	0.159
$\widehat{CE}_{30;K=1}$	0.468	0.198	0.160				

was higher than the conditional effect for male participants ($\widehat{CE}_{20;K=0} = 0.499$, 95% $CI = [-0.249, 1.247]$, $ES = 0.171$).

While the average treatment effect of the speed of processing training was not significant, we found a significant conditional effect for female participants ($\widehat{CE}_{30;K=1} = 0.468$, 95% $CI = [0.080, 0.855]$, $ES = 0.160$). Conversely, for male participants, the speed of processing training did not yield a significant effect ($\widehat{CE}_{30;K=0} = -0.097$, 95% $CI = [-0.830, 0.636]$, $ES = -0.033$). This differential effect illustrates why examining conditional effects can be crucial when evaluating a treatment.

Conditional effects given X and XK

The conditional treatment effects given a treatment condition (and gender) are given in detail in Table 4.2. As the ACTIVE study was a randomized controlled trial, the XK -conditional distributions of depression ξ_1 and baseline test scores ξ_2 should not differ across levels of X . Thus, the conditional effects given a treatment condition are expected to be close to the corresponding average treatment effects. The same applies to conditional effects given treatment and gender. We will not discuss all of these effects in detail here, but rather illustrate their interpretation using one example.

For example, the conditional treatment effect of the memory training for the non-treated ($\widehat{CE}_{10;X=0} = 0.227$, 95% $CI = [-0.114, 0.568]$, $ES = 0.078$) is similar to the average effect of the memory training. This refers to the expected effect if participants who were assigned to the no-contact control group would instead have been assigned to the memory training. These effects being close to the average treatment effect $\widehat{AE}_{10}$ reflects that differences in baseline variables are small across groups. Note that for observational studies or broken experiments, these effects would not be necessarily close to each other due to possible baseline differences.

4.7 Summary and Conclusions

In this paper, we presented and illustrated a new method of accounting for non-normal covariates when estimating average and conditional treatment effects for count outcomes. We extended the moment-based approach by Kiefer and Mayer (2019, *in press*) in three aspects. First, we presented four ways to account for non-normal covariates: (a) for categorical covariates, applying a saturated model; for continuous covariates, we suggested either (b) the use of an alternative, known non-normal joint distribution (e.g., skew-normal distribution), (c) a plausible factorization into marginal and conditional distributions, or (d) approximation via a finite Gaussian mixture distribution. Second, we extended the effect analysis to multiple treatment conditions, making it possible to evaluate the effectiveness of several treatments simultaneously. Third, we introduced conditional effects given a treatment condition and/or values of the categorical covariates into the moment-based approach, allowing for a more fine-grained effect analysis. Finally, we provided an

illustrative example to show how our extensions of the moment-based approach can be applied to real data. In our example, three cognitive training conditions were compared to a no-contact control group regarding the count of correctly answered items in a cognitive reasoning test. We considered the baseline count of correctly answered items (i.e., non-normal covariate), baseline depression (i.e., latent covariate), and gender (i.e., categorical covariate) as covariates. The corresponding negative binomial multigroup structural equation model and the average and conditional treatment effect computations were estimated in R. We made these functions conveniently accessible for applied researchers in an R package.

The aforementioned advancements bring the benefits of the moment-based approach for statistical inference to a broader range of applied scenarios in psychological, social, and health sciences. In contrast to earlier approaches, the moment-based approach treats observed group sizes and covariates' values as random and not predetermined by the experimenter. Ignoring this randomness would lead to underestimation of standard errors and, thus, inflated Type 1 error rates and decreased power, which has previously been shown for stochastic covariates (Li, McLouth, & Delaney, 2019; Y. Liu et al., 2017) and stochastic group sizes (Mayer & Thoemmes, 2019). In addition, the moment-based approach allows to account for measurement error in covariates. For an overview of consequences of measurement error in non-linear regression models, see Carroll et al. (2010). However, in the social and health sciences, observed variables in real datasets often deviate from the normal distribution (Bono et al., 2017; Micceri, 1989). While Kiefer and Mayer (in press) proposed a moment-based approach assuming strictly multivariate normally distributed covariates, we relaxed this assumption in this paper offering several alternatives for non-normally distributed variables. Thus, the moment-based approach with our extension should better fit typical applications in social and health sciences.

4.7.1 Limitations and Further Research

There are some aspects not covered in this article that could provide starting points for further research and refinements of the approach:

In Equation (4.3), we assumed a linear parameterization for the predictor function $f_{xk}(\xi)$. While this is the standard parameterization of, for example, generalized linear models (McCullagh & Nelder, 1996) and for effect analysis with linear link function (e.g., Mayer et al., 2016), sometimes non-linear predictor functions might be of substantial interest. For example, Mayer et al. (2017) investigated the effectiveness of an autonomy support by ninth grade teachers in reducing students' state of boredom. They hypothesized a quadratic relationship between boredom and self-efficacy, which means, the treatment could be most effective for medium values of self-efficacy and less for both extremes. Similarly, Y. Liu and West (2015) use trigonometric terms to predict cyclic patterns in daily report (count) data. The inclusion of quadratic or other non-linear terms might also be interesting for count outcomes and regression models with a logarithmic link function. We

presume, that the moment-based approach would not yield analytical effect formulas for non-linear predictor functions, but effect computation using numerical integration would be an alternative.

Throughout the paper, we factorized the joint distribution of the covariates into a marginal distribution of the categorical variables (i.e., treatment X and covariates K) and a XK -conditional distribution of the continuous covariates. Our choice can seem restrictive, as alternative factorizations might be more suitable for some cases. For example, [Kiefer and Mayer \(2019\)](#) used a converse factorization, that is, a marginal distributed continuous covariate Z and a Z -conditional probability of the binary treatment X . However, we would argue that for a given joint distribution multiple equivalent factorizations exist. In the aforementioned example, the joint distribution can be equivalently re-written in our factorization, that is, a marginal distribution of the binary treatment and X -conditional distribution of Z (for a derivation, see [Appendix 4.A](#)).

We used maximum likelihood estimation to obtain parameter estimates for the negative binomial multigroup structural equation model, but in applied settings alternative statistical frameworks might be more suitable. Maximum likelihood estimation can suffer from non-convergence issues in complex models ([Deng, Yang, & Marcoulides, 2018](#)). We also exhibited this phenomenon in our illustrative example, where we had to exclude overdispersion parameters from the model in order to achieve convergence. Thus, we recommend three solutions to approach non-convergence issues. First, a pragmatic work-around would be to specify a more parsimonious model, because non-convergence might be a consequence of overparameterization ([Jackson, 2001, 2003](#)). Second, a Bayesian structural equation modeling framework can be applied. The Bayesian approach can typically handle complex models much easier than MLE ([Merkle & Rosseel, 2015](#)), even such models which are nonidentified in MLE ([Muthén & Asparouhov, 2012](#)). Third, recent factor score approaches (e.g., [Devlieger & Rosseel, 2017](#)) might also help achieving convergence by simplifying the measurement model and, thereby, the numerical integration part of the MLE.

The aforementioned solutions for estimating more complex models may become even more important if further practical complications of the conducted experiment (e.g., treatment non-compliance) or data quality (e.g., missing data) should be accounted for. For missing data handling, full information maximum likelihood (FIML; [Arbuckle, 1996](#)) is a viable option within the structural equation modeling framework. However, multiple imputation ([Enders, 2010](#)) might be a reasonable alternative in some cases, because it spares the necessity to introduce additional numerical integration to the log-likelihood function. In cases of *broken experiments* ([Barnard, Frangakis, Hill, & Rubin, 2003](#)), additional elaborate modeling approaches might be required. For example, if participants do not comply with their assigned treatment condition (e.g., assigned to cognitive training, but not visiting (all) training sessions), [Sagarin et al. \(2014\)](#) propose several techniques, for example, to estimate the complier average causal effect or to use dose-response models.

In recent years, multilevel versions of count regression models have been embraced

(Aiken, Mistler, Cox, & West, 2015; Austin, Stryhn, Leckie, & Merlo, 2017; McNeish, 2019) for datasets where participants are nested in clusters, for example, students in classes. Employing a multilevel framework, both sampling dependencies among participants and contextual covariates can be included in the effect analysis. Mayer, Nagengast, et al. (2014) provide a framework of average and conditional treatment effects using multilevel structural equation modeling. The traditional two-level SEM approach (Longford & Muthén, 1992; Muthén, 1994) is based on a decomposition into a within- and a between-level SEM. If count variables are involved, however, this approach cannot be straightforwardly adapted, because a between-within decomposition of a count variable would result in two *non-count* variables. Alternatively, an average and conditional treatment effect framework based on generalized structural equation models (Rabe-Hesketh, Skrondal, & Pickles, 2004), which do not require a between-within decomposition, could be developed.

In our illustrative example, we examined the effectiveness of cognitive trainings using a pre-post-design. The ACTIVE study also contains several follow-up measurements allowing to investigate long-term effectiveness of the cognitive trainings. In future developments, it might be fruitful to incorporate this longitudinal information into effect analyses. We suggest two ways for doing so. First, when examining long-term effects using a follow-up measurement, earlier measurements (e.g., post-test) could be included as mediators. In this scenario, total, direct, and indirect effects of the treatment can be distinguished (Mayer, Thoemmes, et al., 2014). This can help researchers understand, for example, which post-test characteristics foster a long-term effect of the cognitive training. Second, several measurement occasions can be summarized or contrasted, for example, using a growth curve model (McArdle & Epstein, 1987) or growth component model (Kiefer, Rosseel, Wiese, & Mayer, 2018; Mayer, Steyer, & Mueller, 2012) among follow-up measurements. In this case, the outcome-of-interest would be a latent slope or growth component instead of the count outcome.

In nonrandomized observational studies the treatment effects discussed in this paper are not necessarily causal effects. However, causality theories, such as Rubin's causal model (Rubin, 2005) or the stochastic theory of causal effects (Steyer et al., 2014) provide causality conditions under which causal effects can be estimated. Usually, these conditions require a careful selection of covariates, which then are controlled for with regression-adjustment (e.g., Mayer, 2019; Mayer et al., 2016) or propensity score methods (Rosenbaum & Rubin, 1983). For an overview of design and analysis in quasi-experimental settings, see also Reichardt (2019).

Appendix

4.A Factorization of the Joint Distribution

In Equations (4.4) and (4.5), we claimed that our factorization of the joint distribution was without loss of generality. This can be shown using two properties of distribution functions. Let us consider two random variables A and B with joint distribution $p_{A,B}(a, b)$, then

$$p_{A,B}(a, b) = p_A(a) \cdot p_{B|A}(b|a) = p_B(b) \cdot p_{A|B}(a|b)$$

$$p_A(a) = \int_b p_{A,B}(a, b) db.$$

If the marginal distribution $p_B(b)$ and the conditional distribution $p_{A|B}(a|b)$ are known, the corresponding distributions $p_A(a)$ and $p_{B|A}(b|a)$ can directly be computed with

$$p_A(a) = \int_b p_B(b) \cdot p_{A|B}(a|b) db$$

$$p_{B|A}(b|a) = \frac{p_B(b) \cdot p_{A|B}(a|b)}{p_A(a)} = \frac{p_B(b) \cdot p_{A|B}(a|b)}{\int_b p_B(b) \cdot p_{A|B}(a|b) db}$$

Consequently, if one version or factorization of the joint distribution is known, equivalent factorizations for the same joint distribution can be derived.

Let us now consider a simple example of the moment based approach, with a single covariate Z and a binary treatment variable X . In contrast to our factorization in Equations (4.4) and (4.5), we now consider

$$p_Z(z) \cdot P(X = x|Z = z)$$

as known factorization of the joint distribution, where $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$ and the conditional probability of $X = x$ is given by

$$P(X = x|Z = z) = \left[\frac{\exp(v_0 + v_1 z)}{1 + \exp(v_0 + v_1 z)} \right]^x \cdot \left[1 - \frac{\exp(v_0 + v_1 z)}{1 + \exp(v_0 + v_1 z)} \right]^{(1-x)}$$

that is, as a logistic function of the covariate Z . This kind of joint distribution can be useful if self selection of participants into treatment $X = 1$ or control $X = 0$ depending on the covariate Z is assumed.

Nevertheless, our factorization from Equation (4.5) can be obtained from these known marginal and conditional distributions by computing

$$p_X(x) = \int_z p_Z(z) \cdot P(X = x|Z = z) dz$$

$$p_{Z|X} = \frac{p_Z(z) \cdot P(X = x|Z = z)}{\int_z p_Z(z) \cdot P(X = x|Z = z) dz}$$

Note: These transformations are generally required for effect computation only. The maximum likelihood estimation would, nevertheless, be based on $p_Z(z)$ and $P(X = x|Z = z)$ in this example. In contrast, when trying to estimate the likelihood function based on $P(X = x)$ and $p_{Z|X}$, it might be tempting to estimate $P(X = x)$ directly instead of $\int_z p_Z(z) \cdot P(X = x|Z = z) dz$. However, introducing a parameterization for $P(X = x)$ would lead to redundancies, because $P(X = x)$ is already determined by μ_Z , σ_Z^2 , v_0 , and v_1 . Consequently, additional estimation of $P(X = x)$ would lead to a non-invertible covariance matrix of the parameter estimates and, thus, standard errors could not be computed. The redundant parameter would remain on its starting value and, thus, would not yield a trustworthy estimate. The likelihood function should only contain parameters μ_Z , σ_Z^2 , v_0 , and v_1 . Corresponding parameters from our factorization and their standard errors can then be computed for effect estimation using the delta method.

4.B Model Estimation

In this section, we present details on the maximum likelihood estimation of the NB-MG-SEM for our illustrative example. Let us consider a count outcome variable Y , a treatment variable X with 4 levels, a binary categorical covariate K , and a vector of continuous covariates $\boldsymbol{\xi} = (\xi_1, \xi_2)$, containing a latent covariate ξ_1 measured by three observed indicator variables $\mathbf{z} = (Z_1, Z_2, Z_3)$, and a count covariate ξ_2 . We want to estimate a statistical model as presented in Equations (4.7), (4.8), and (4.9) using maximum-likelihood techniques. Assuming a sample of N independently and identically distributed observations, the complete data likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}|Y, X, K, \mathbf{z}, \boldsymbol{\xi}) = \prod_{i=1}^N f(y_i, x_i, k_i, \mathbf{z}_i, \boldsymbol{\xi}_i|\boldsymbol{\theta})$$

where $y_i, x_i, k_i, \mathbf{z}_i, \boldsymbol{\xi}_i$ are the i -th observed values of the aforementioned variables. As we do not have observed values ξ_1 (i.e., latent covariate), we actually have to optimize the observed data likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|Y, X, K, \mathbf{z}, \xi_2) &= \prod_{i=1}^N f(y_i, x_i, k_i, \mathbf{z}_i, \xi_{2i}|\boldsymbol{\theta}) \\ &= \prod_{i=1}^N \int_{\xi_1} f(y_i, x_i, k_i, \mathbf{z}_i, \boldsymbol{\xi}_i|\boldsymbol{\theta}) d\xi_1 \end{aligned}$$

Note that marginalizing over the joint density of the complete data is only one way to optimize this likelihood. It would also be possible to use an EM algorithm (Dempster et al., 1977) to solve the complete data likelihood, but this would be beyond the scope of this manuscript.

The joint density $f(y_i, x_i, k_i, z_i, \xi_i | \boldsymbol{\theta})$ can be factorized into five conditional densities:

$$\begin{aligned} f(y_i, x_i, k_i, z_i, \xi_i | \boldsymbol{\theta}) = & f(n_{x_i k_i} | \boldsymbol{\theta}_1) \\ & \cdot f(y_i | \boldsymbol{\theta}_2, x_i, k_i, \xi_i) \\ & \cdot f(z_i | \boldsymbol{\theta}_3, x_i, k_i, \xi_{1i}) \\ & \cdot f(\xi_{2i} | \boldsymbol{\theta}_4, x_i, k_i, \xi_{1i}) \\ & \cdot f(\xi_{1i} | \boldsymbol{\theta}_5, x_i, k_i) \end{aligned}$$

In the last step, we decomposed the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4, \boldsymbol{\theta}_5)$ into four parts corresponding to the four conditional densities. The parameters to estimate in the respective densities are:

$$\begin{aligned} \boldsymbol{\theta}_1 &= (\kappa_{00}, \dots, \kappa_{pj}) \\ \boldsymbol{\theta}_2 &= (\alpha_{00}, \dots, \alpha_{pj}) \\ \boldsymbol{\theta}_3 &= (\boldsymbol{\nu}, \boldsymbol{\Lambda}, \boldsymbol{\Theta}_{00}, \dots, \boldsymbol{\Theta}_{pj}) \\ \boldsymbol{\theta}_4 &= (\gamma_{00}, \dots, \gamma_{pj}) \\ \boldsymbol{\theta}_5 &= (\mu_{00}, \dots, \mu_{pj}, \sigma_{00}^2, \dots, \sigma_{pj}^2) \end{aligned}$$

In the NB-MG-SEM, the conditional densities are defined by the model assumptions. Here, $f(n_{x_i k_i} | \boldsymbol{\theta}_1)$ reflects the model for group sizes n_{xk} , which are assumed to be Poisson distributed. The group-specific Poisson regression of Y on $\boldsymbol{\xi}$ is given by $f(y_i | \boldsymbol{\theta}_2, x_i, k_i, \xi_i)$, the group-specific and $\xi_1 = \xi_{1i}$ -conditional distribution of the observed indicators $\mathbf{z}$ is given by $f(z_i | \boldsymbol{\theta}_3, x_i, k_i, \xi_{1i})$, the group-specific Poisson regression of ξ_2 on ξ_1 is given by $f(\xi_{2i} | \boldsymbol{\theta}_4, x_i, k_i, \xi_{1i})$, and the group-specific marginal distribution of the latent variable ξ_1 is given by $f(\xi_{1i} | \boldsymbol{\theta}_5, x_i, k_i)$. Thus, the conditional densities in our case are

$$\begin{aligned} f(n_{x_i k_i} | \boldsymbol{\theta}_1) &= \frac{\exp(\kappa_{xk}^{n_{x_i k_i}})}{n_{x_i k_i}!} \exp(-\exp[\kappa_{xk}]) \\ f(y_i | \boldsymbol{\theta}_2, x_i, k_i, \xi_i) &= \frac{\exp(\alpha'_{xk} \xi_i)^{y_i}}{y_i!} \exp(-\exp[\alpha'_{xk} \xi_i]) \\ f(z_i | \boldsymbol{\theta}_3, x_i, k_i, \xi_{1i}) &= \frac{1}{\sqrt{(2\pi)^3 |\boldsymbol{\Theta}_{xk}|}} \exp\left(-\frac{1}{2}(\mathbf{z}_i - \xi_{1i})' \boldsymbol{\Theta}_{xk}^{-1} (\mathbf{z}_i - \xi_{1i})\right) \\ f(\xi_{2i} | \boldsymbol{\theta}_4, x_i, k_i, \xi_{1i}) &= \frac{\exp(\gamma'_{xk}(1, \xi_{1i})')^{\xi_{2i}}}{\xi_{2i}!} \exp(-\exp[\gamma'_{xk}(1, \xi_{1i})']) \\ f(\xi_{1i} | \boldsymbol{\theta}_5, x_i, k_i) &= \frac{1}{\sqrt{(2\pi\sigma_{xk}^2)}} \exp\left(-\frac{(\xi_{1i} - \mu_{xk})^2}{2\sigma_{xk}^2}\right) \end{aligned}$$

As the group weights do not depend on the latent covariate ξ_2 , we can simplify the

observed data density as follows:

$$\begin{aligned} f(y_i, x_i, k_i, \mathbf{z}_i, \xi_{2i} | \boldsymbol{\theta}) &= \int_{\xi_1} f(y_i, x_i, k_i, \mathbf{z}_i, \xi_i | \boldsymbol{\theta}) d\xi_1 \\ &= f(n_{x_i k_i} | \boldsymbol{\theta}) \cdot \int_{\xi_1} f(y_i, \mathbf{z}_i, \boldsymbol{\xi} | \boldsymbol{\theta}, x_i, k_i) d\xi_1 \end{aligned}$$

Hence, the corresponding log-likelihood is

$$\log \mathcal{L}(\boldsymbol{\theta} | Y, X, K, \mathbf{z}, \boldsymbol{\xi}) = \sum_{i=1}^N \log[f(n_{x_i k_i} | \boldsymbol{\theta})] + \sum_{i=1}^N \log \left[\int_{\xi_1} f(y_i, \mathbf{z}_i, \boldsymbol{\xi} | \boldsymbol{\theta}, x_i, k_i) d\xi_1 \right]$$

As there is no closed form solution for $\int_{\xi_1} f(y_i, \mathbf{z}_i, \boldsymbol{\xi} | \boldsymbol{\theta}, x_i, k_i) d\xi_1$, numerical integration is required. In our implementation, we use Gauss-Hermite quadrature which approximates the integral over ξ_1 with a finite sum

$$\begin{aligned} &\int_{\xi_1} f(y_i, \mathbf{z}_i, \boldsymbol{\xi} | \boldsymbol{\theta}, x_i, k_i) d\xi_1 \\ &\approx \sum_{h=1}^H w_h \cdot f(y_i | \boldsymbol{\theta}_2, x_i, k_i, \xi_{1h}^*, \xi_{2i}) \cdot f(\mathbf{z}_i | \boldsymbol{\theta}_3, x_i, k_i, \xi_{1h}^*) \cdot f(\xi_{2i} | \boldsymbol{\theta}_4, x_i, k_i, \xi_{1h}^*) \end{aligned}$$

where $\xi_{1h}^* = \sqrt{2\sigma_{xk}^2} a_h + \mu_{xk}$ and a_h, w_h are Gauss-Hermite quadrature points and weights.

4.C Effect Estimation

In this section, we provide details on the derivation of the average and conditional effect formulas used in our illustrative example. Let us consider a count outcome variable Y , a treatment variable X with 4 levels, a binary categorical covariate K , and a vector of continuous covariates $\boldsymbol{\xi} = (\xi_1, \xi_2)$, containing a latent covariate ξ_1 measured by three observed indicator variables $\mathbf{z} = (Z_1, Z_2, Z_3)$, and a count covariate ξ_2 . In the model section, we already stated that we assume latent baseline depression is XK -conditionally marginally normally distributed $\xi_1 | X = x, K = k \sim \mathcal{N}(\mu_{xk}, \sigma_{xk})$ and is related to the baseline count of correctly answered items via a Poisson regression with $\xi_2 | X = x, K = k, \xi_1 \sim \mathcal{P}(\lambda_{xk})$ with $\lambda_{xk} = \exp(\gamma_{xk0} + \gamma_{xk1}\xi_1)$.

Beginning with the general definition of the average treatment effect for these variables

$$\begin{aligned} ATE &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\ &\quad \int_{\xi_1} \int_{\xi_2} (\exp[\boldsymbol{\alpha}'_{tk} \boldsymbol{\xi}] - \exp[\boldsymbol{\alpha}'_{0k} \boldsymbol{\xi}]) \cdot p_{\xi_1 | K, X}(\xi_1 | k, x) p_{\xi_2 | \xi_1, K, X}(\xi_2 | \xi_1, k, x) d\xi_1 d\xi_2 \end{aligned}$$

the integration part can be rewritten as a difference of integrals

$$\begin{aligned}
 &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\
 &\quad \left(\int_{\xi_1} \int_{\xi_2} \exp[\alpha'_{tk} \xi] \cdot p_{\xi_1|K,X}(\xi_1|k, x) p_{\xi_2|\xi_1, K, X}(\xi_2|\xi_1, k, x) d\xi_1 d\xi_2 \right. \\
 &\quad \left. - \int_{\xi_1} \int_{\xi_2} \exp[\alpha'_{0k} \xi] \cdot p_{\xi_1|K,X}(\xi_1|k, x) p_{\xi_2|\xi_1, K, X}(\xi_2|\xi_1, k, x) d\xi_1 d\xi_2 \right)
 \end{aligned}$$

where the decomposition of the joint distribution of ξ_1 and ξ_2 into a marginal and a conditional distribution enables the formation of an inner and an outer integral

$$\begin{aligned}
 &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\
 &\quad \left(\int_{\xi_1} \exp[\alpha_{tk0} + \alpha_{tk1} \xi_1] \left[\int_{\xi_2} \exp[\alpha_{tk2} \xi_2] \cdot p_{\xi_2|\xi_1, K, X}(\xi_2|\xi_1, k, x) d\xi_2 \right] p_{\xi_1|K, X}(\xi_1|k, x) d\xi_1 \right. \\
 &\quad \left. - \int_{\xi_1} \exp[\alpha_{0k0} + \alpha_{0k1} \xi_1] \left[\int_{\xi_2} \exp[\alpha_{0k2} \xi_2] \cdot p_{\xi_2|\xi_1, K, X}(\xi_2|\xi_1, k, x) d\xi_2 \right] p_{\xi_1|K, X}(\xi_1|k, x) d\xi_1 \right)
 \end{aligned}$$

The inner integral resembles a univariate (conditional) expectation and can therefore be substituted with a univariate moment-generating function:

$$\begin{aligned}
 \int_{\xi_2} \exp[\alpha_{xk2} \xi_2] \cdot p_{\xi_2|\xi_1, K, X}(\xi_2|\xi_1, k, x) d\xi_2 &= E(\exp[\alpha_{xk2} \xi_2] | X, K, \xi_1) \\
 &= M_{\xi_2|\xi_1, K, X}(\alpha_{xk2})
 \end{aligned}$$

Consequently, the computation of the average treatment effect simplifies to

$$\begin{aligned}
 ATE &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\
 &\quad \left(\int_{\xi_1} \exp[\alpha_{tk0} + \alpha_{tk1} \xi_1] [M_{\xi_2|\xi_1, K, X}(\alpha_{tk2})] p_{\xi_1|K, X}(\xi_1|k, x) d\xi_1 \right. \\
 &\quad \left. - \int_{\xi_1} \exp[\alpha_{0k0} + \alpha_{0k1} \xi_1] [M_{\xi_2|\xi_1, K, X}(\alpha_{0k2})] p_{\xi_1|K, X}(\xi_1|k, x) d\xi_1 \right)
 \end{aligned}$$

and as both outer integrals have the same domain, we can go back to a single integral

$$\begin{aligned}
 &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\
 &\quad \int_{\xi_1} \left[\exp[\alpha_{tk0} + \alpha_{tk1} \xi_1] M_{\xi_2|\xi_1, K, X}(\alpha_{tk2}) \right. \\
 &\quad \left. - \exp[\alpha_{0k0} + \alpha_{0k1} \xi_1] M_{\xi_2|\xi_1, K, X}(\alpha_{0k2}) \right] p_{\xi_1|K, X}(\xi_1|k, x) d\xi_1
 \end{aligned}$$

where we integrate over the density of the XK -conditionally normally distributed ξ_1 ,

which allows us to approximate the integral using a Gauss-Hermite quadrature

$$\begin{aligned}
 &= \sum_{x=0}^p \sum_{k=0}^j P(X = x, K = k) \cdot \\
 &\quad \sum_{h=1}^H w_h \left[\exp(\alpha_{tk0} + \alpha_{tk1} \xi_{1h}^*) M_{\xi_2|\xi_1, K, X}(\alpha_{tk2}) - \exp(\alpha_{0k0} + \alpha_{0k1} \xi_{1h}^*) M_{\xi_2|\xi_1, K, X}(\alpha_{0k2}) \right]
 \end{aligned}$$

where $\xi_{1h}^* = \sqrt{2\sigma_{xk}^2} a_h + \mu_{xk}$ and a_h, w_h are Gauss-Hermite quadrature points and weights.

4.D Supplementary Material: R Code for Model Estimation

```

1 #####
2 # load required packages
3 #####
4 library(MplusAutomation)
5 library(pracma)
6 library(lavaan)
7 library(dplyr)
8 library(data.table)
9
10 #####
11 # readData
12 #####
13 d <- readRDS("dACTIVE.rds")
14 dNA <- na.omit(d)
15 N <- nrow(dNA)
16
17 #####
18 ### Pre-Building extended data structure for GH quadrature
19 #####
20 intPoints <- 21 # number of integration points
21 ghPoints <- pracma::gaussHermite(intPoints) # get GH weights and integration points
22
23 # preparing additional variables for memory allocation
24 mu.y1 <- mu.y2 <- mu.y3 <- rep(0, intPoints)
25 f.y1 <- f.y2 <- f.y3 <- rep(0, intPoints)
26 mu.z1 <- f.z1 <- LAMBDA <- f.counts <- rep(0, intPoints)
27
28 # gather GH quadrature points and weights and additional variables
29 extenDF <- data.frame(ghEta = ghPoints$x,
30 ghEtaFix = ghPoints$x,
31 ghWeights = ghPoints$w,
32 mu.y1, mu.y2, mu.y3,
33 f.y1, f.y2, f.y3,
34 mu.z1, f.z1,
35 LAMBDA, f.counts)
36 id <- rep(c(1:N), each = intPoints)
37
38 ## extend original data frame with GH quadrature information
39 Data <- cbind(ID = as.factor(id), dNA[id,], extenDF, out = NA)
40
41 #####
42 # Start values
43 #####
44 x.start.mm <- c(1, 1, # lambda2,3
45 1, 1, 1, # theta1,2,3
46 0, 0) # nu2,3
47 x.start.xk <- c(0, 1, #mu, psi in cell
48 0, 0.1, # regression coefs count on latent
49 0, 0.1, 0.1) # regression coefs y on count and latent
50 x.start.kappa <- c(rep(log(N/8),8)) # 7 kappas for 8 cells
51 x.start <- c(x.start.mm,
52 rep(x.start.xk, 8))
53
54 #####
55 ### Log-likelihood functiond #
56 #####
57 loglik <- function(x) {
58 # No missing values in the parameter vector

```

```
59   if(anyNA(x)){return(+Inf)}
60
61   ## map 'x' to parameter matrices
62   # invariant parameters
63   # factor loadings
64   lambda1 <- 1; lambda2 <- x[1]; lambda3 <- x[2]
65   # residual variances
66   theta1 <- x[3]; theta2 <- x[4]; theta3 <- x[5]
67   # factor intercepts
68   nu1 <- 0; nu2 <- x[6]; nu3 <- x[7]
69
70   # XK-specific parameters (mu, sigma, gammas, alphas)
71   xk.mat <- x[8:63] %>% matrix(ncol = 7, byrow = T)
72   colnames(xk.mat) <- c("mu", "psi", "g0", "g1", "a0", "a1", "a2")
73   xk.mat <- as.data.frame(xk.mat)
74
75   # no negative variances...
76   if(theta1 <= 0 || theta2 <= 0 || theta3 <= 0 || any(xk.mat$psi <= 0) ) return(+Inf)
77
78   # compute xi_1* values of the Gauss-Hermite quadrature
79   Data$ghEta <- sqrt(2*xk.mat$psi[Data$cell])*Data$ghEtaFix + xk.mat$mu[Data$cell]
80
81   # expectation of z1/z2/z3 for this given value of xi_1*
82   Data$mu.y1 <- nu1 + lambda1*Data$ghEta
83   Data$mu.y2 <- nu2 + lambda2*Data$ghEta
84   Data$mu.y3 <- nu3 + lambda3*Data$ghEta
85
86   # f(z/xi); likelihood of observed indicators
87   Data$f.y1 <- dnorm(Data$ind1, mean = Data$mu.y1, sd = sqrt(theta1), log = FALSE)
88   Data$f.y2 <- dnorm(Data$ind2, mean = Data$mu.y2, sd = sqrt(theta2), log = FALSE)
89   Data$f.y3 <- dnorm(Data$ind3, mean = Data$mu.y3, sd = sqrt(theta3), log = FALSE)
90
91   # f(xi_2/x_1) as Poisson regression
92   Data$mu.z1 <- exp(xk.mat$g0[Data$cell] + xk.mat$g1[Data$cell]*Data$ghEta)
93   Data$f.z1 <- dpois(Data$pretest, lambda = Data$mu.z1, log = FALSE)
94
95   # lik counts Y for this given value of xi
96   Data$LAMBDA <- exp(xk.mat$a0[Data$cell] + xk.mat$a1[Data$cell]*Data$ghEta + xk.mat$a2[
     Data$cell]*Data$pretest)
97   Data$f.counts <- dpois(Data$dv, lambda = Data$LAMBDA, log = FALSE)
98
99   # summands of Gauss-Hermite quadrature
100  Data$out <- Data$ghWeights*Data$f.z1*Data$f.y1*Data$f.y2*Data$f.y3*Data$f.counts/sqrt(pi
    )
101
102  # summing of Gauss-Hermite quadrature summands
103  tmp <- setDT(Data)[,(x = sum(out)), by=ID]
104  if(anyNA(tmp)) return(+Inf)
105
106  # computing and summing over the individual log-likelihood
107  out <- tmp$x %>% log() %>% sum()
108
109  # rescaling for nlminb
110  obj <- -1 * out
111
112  # rescale to make nlminb happy
113  obj <- obj / (nrow(Data)/intPoints)
114
115  obj
116  }
117
```

```
118 # log-likelihood for group sizes
119 loglik_group <- function(x){
120 # stochastic group size
121 expkappas <- exp(x)
122 n.cell <- table(Data$cell)/intPoints
123
124 # Poisson group model
125 obj_group <- sum(-expkappas + n.cell*log(expkappas) - lgamma(n.cell + 1))
126 obj <- - obj_group/(nrow(Data)/intPoints)
127
128 obj
129 }
130
131 #####
132 # Optimization of log-likelihood
133 #####
134 ## Optimize log-likelihood for non-group part
135 out <- nlminb(start = x.start,
136 objective = loglik,
137 control = list(rel.tol = 1e-6)
138 )
139
140 ## Optimize log-likelihood for group part
141 out_group <- nlminb(start = x.start.kappa,
142 objective = loglik_group,
143 control = list(rel.tol = 1e-6)
144 )
145
146 par_final <- c(out$par, out_group$par)
147
148 ## Standard Errors for Model (via numerical derivation)
149 # for non-group part
150 H <- pracma::hessian(loglik, out$par) # may take a while...
151 I <- H # Note: rescaling (-1* and /N) has already taken place in the loglik-function
152 varcov <- solve(I)/N
153
154 # for group part
155 H_group <- pracma::hessian(loglik_group, out_group$par)
156 I_group <- H_group
157 varcov_group <- solve(I_group)/N
158
159 # join both vcovs (parameters from group and non-group part are independent)
160 vcov_final <- lavaan::lav_matrix_bdiag(varcov, varcov_group)
```

4.E Supplementary Material: R Code for Effect Estimation

```
1 library(dplyr)
2 library(data.table)
3 library(lavaan)
4
5 #####
6 # required: par_final and vcov_final from model_estimation
7 #####
8
9 #### Effect Estimation ####
10 computeCondEffect <- function(x){
11 ## This function computes conditional treatment effects within the XK-groups
12 ## i.e., only the integration part given any combination of x and k
13 ## aggregation to ATE and CE is done in another function
```

```
14
15 # XK-specific parameters
16 xk.mat <- x[8:63] %>% matrix(ncol = 7, byrow = T)
17 xk.mat <- cbind(xk.mat, c(0,1), c(0,0,1,1,2,2,3,3))
18 colnames(xk.mat) <- c("mu", "psi", "g0", "g1", "a0", "a1", "a2", "k", "x")
19
20 # stochastic group size
21 kappas <- x[64:71]
22
23 # re-arranging the parameters for effect estimation
24 gFuncs <- data.frame(xk.mat[,c("a0", "a1", "a2", "k", "x")])
25 covFuncs <- data.frame(xk.mat[,c("mu", "psi", "g0", "g1", "x", "k")])
26 conditions <- expand.grid(k=c(0,1),x=c(0,1,2,3),gx=c(1,2,3), eff =NA)
27
28 ## GH quadrature init
29 intPoints <- 20
30 ghPoints <- pracma::gaussHermite(intPoints)
31
32 ## Effect for treatments t in any combination of x and k
33 for (i in 1:nrow(conditions)){
34   k <- conditions$k[i]
35   x <- conditions$x[i]
36   gx <- conditions$gx[i]
37
38   # parameters of g-function
39   A1 <- gFuncs[gFuncs$x == gx & gFuncs$k == k, c("a0", "a1", "a2")]
40   A0 <- gFuncs[gFuncs$x == 0 & gFuncs$k == k, c("a0", "a1", "a2")]
41
42   # parameters of covariates
43   covPar <- covFuncs[covFuncs$x == x & covFuncs$k == k, ]
44
45   # group-specific values of xi_1*
46   eta <- sqrt(2*covPar$psi)*ghPoints$x + covPar$mu
47
48   # Poisson regression of xi_2 on xi_1
49   muPre <- exp(covPar$g0 + covPar$g1*eta)
50
51   # moment-generating functions for xi_2
52   mgf1 <- exp((muPre)*(exp(A1$a2) - 1))
53   mgf0 <- exp((muPre)*(exp(A0$a2) - 1))
54
55   # summands and summation of GH quadrature
56   out <- ( exp(A1$a0 + A1$a1*eta)*mgf1 - exp(A0$a0 + A0$a1*eta)*mgf0 )*ghPoints$w/sqrt(pi)
57   conditions$eff[i] <- sum(out)
58 }
59
60 # computing the probabilities P(X=x,K=k)
61 Pk <- exp(kappas)/(sum(exp(kappas)))
62 return(c(conditions$eff, Pk))
63 }
64 # apply basic effect estimation function to estimated parameters
65 ce <- computeCondEffect(par_final)
66
67
68 # standard error for CondEffect function
69 # computes variance of xk-conditional effects and group probabilities (transformations
   of model parameters)
70 # delta method
71 delta.method.ck <- function(func, x, varcov){
72   func.jacobian <- lav_func_jacobian_complex(func, x)
73   func.jacobian %*% varcov %*% t(func.jacobian)
```

```

74 }
75 vcov.ce <- delta.method.ck(computeCondEffect, par_final, varcov = vcov_final)
76 vcov.ce <- lav_matrix_bdiag(vcov.ce[1:24,1:24], vcov.ce[25:32,25:32])
77 ce.se <- sqrt(diag(vcov.ce))
78
79
80 #####
81 ##### Lavaan syntax for aggregated effects
82 lsyn <- '
83 # Definition of free parameters re. conditional effects
84 ## Note: These "free" parameters are plugged in from our own model estimation
85 g1 ~ CE100*xk00 + CE101*xk01 + CE110*xk10 + CE111*xk11 + CE120*xk20 + CE121*xk21 + CE130
      *xk30 + CE131*xk31
86 g2 ~ CE200*xk00 + CE201*xk01 + CE210*xk10 + CE211*xk11 + CE220*xk20 + CE221*xk21 + CE230
      *xk30 + CE231*xk31
87 g3 ~ CE300*xk00 + CE301*xk01 + CE310*xk10 + CE311*xk11 + CE320*xk20 + CE321*xk21 + CE330
      *xk30 + CE331*xk31
88 p1 ~ relfreq00*xk00 + relfreq01*xk01 + relfreq10*xk10 + relfreq11*xk11 + relfreq20*xk20
      + relfreq21*xk21 + relfreq30*xk30 + relfreq31*xk31
89
90 ## Unconditional Probabilities P(K=k)
91 Pk0 := relfreq00 + relfreq01 + relfreq10 + relfreq30
92 Pk1 := relfreq01 + relfreq11 + relfreq21 + relfreq31
93
94 ## Unconditional Probabilities P(X=x)
95 Px0 := relfreq00 + relfreq01
96 Px1 := relfreq10 + relfreq11
97 Px2 := relfreq20 + relfreq21
98 Px3 := relfreq30 + relfreq31
99
100 ## Conditional Probabilities P(K=k|X=x)
101 Pk0gx0 := relfreq00/Px0
102 Pk1gx0 := relfreq01/Px0
103 Pk0gx1 := relfreq10/Px1
104 Pk1gx1 := relfreq11/Px1
105 Pk0gx2 := relfreq20/Px2
106 Pk1gx2 := relfreq21/Px2
107 Pk0gx3 := relfreq30/Px3
108 Pk1gx3 := relfreq31/Px3
109
110 ## Conditional Probabilities P(X=x|K=k)
111 Px0gk0 := relfreq00/Pk0
112 Px1gk0 := relfreq10/Pk0
113 Px2gk0 := relfreq20/Pk0
114 Px3gk0 := relfreq30/Pk0
115 Px0gk1 := relfreq01/Pk1
116 Px1gk1 := relfreq11/Pk1
117 Px2gk1 := relfreq21/Pk1
118 Px3gk1 := relfreq31/Pk1
119
120 # Average treatment effects
121 ave10 := CE100*relfreq00 + CE101*relfreq01 + CE110*relfreq10 + CE111*relfreq11 + CE120*
      relfreq20 + CE121*relfreq21 + CE130*relfreq30 + CE131*relfreq31
122 ave20 := CE200*relfreq00 + CE201*relfreq01 + CE210*relfreq10 + CE211*relfreq11 + CE220*
      relfreq20 + CE221*relfreq21 + CE230*relfreq30 + CE231*relfreq31
123 ave30 := CE300*relfreq00 + CE301*relfreq01 + CE310*relfreq10 + CE311*relfreq11 + CE320*
      relfreq20 + CE321*relfreq21 + CE330*relfreq30 + CE331*relfreq31
124
125 # K=k conditional effects
126 CE10gk0 := Px0gk0*CE100 + Px1gk0*CE110 + Px2gk0*CE120 + Px3gk0*CE130
127 CE20gk0 := Px0gk0*CE200 + Px1gk0*CE210 + Px2gk0*CE220 + Px3gk0*CE230

```

```

128 CE30gk0 := Px0gk0*CE300 + Px1gk0*CE310 + Px2gk0*CE320 + Px3gk0*CE330
129 CE10gk1 := Px0gk1*CE101 + Px1gk1*CE111 + Px2gk1*CE121 + Px3gk1*CE131
130 CE20gk1 := Px0gk1*CE201 + Px1gk1*CE211 + Px2gk1*CE221 + Px3gk1*CE231
131 CE30gk1 := Px0gk1*CE301 + Px1gk1*CE311 + Px2gk1*CE321 + Px3gk1*CE331
132
133 # X=x conditional effects
134 CE10gx0 := Pk0gx0*CE100 + Pk1gx0*CE101
135 CE20gx0 := Pk0gx0*CE200 + Pk1gx0*CE201
136 CE30gx0 := Pk0gx0*CE300 + Pk1gx0*CE301
137
138 CE10gx1 := Pk0gx1*CE110 + Pk1gx1*CE111
139 CE20gx1 := Pk0gx1*CE210 + Pk1gx1*CE211
140 CE30gx1 := Pk0gx1*CE310 + Pk1gx1*CE311
141
142 CE10gx2 := Pk0gx2*CE120 + Pk1gx1*CE121
143 CE20gx2 := Pk0gx2*CE220 + Pk1gx1*CE221
144 CE30gx2 := Pk0gx2*CE320 + Pk1gx1*CE321
145
146 CE10gx3 := Pk0gx3*CE130 + Pk1gx3*CE131
147 CE20gx3 := Pk0gx3*CE230 + Pk1gx3*CE231
148 CE30gx3 := Pk0gx3*CE330 + Pk1gx3*CE331
149 '
150 # lavaanify aggregation syntax (create partable)
151 ## computes aggregated effect and corresponding standard errors
152 pt <- lavaanify(lsyn)
153 def.function <- lav_partable_constraints_def(pt)
154 JAC <- lav_func_jacobian_complex(func=def.function, x=ce)
155 info.r <- JAC %*% vcov.ce %*% t(JAC)
156
157 # Extract the point estimates and standard errors
158 est <- def.function(ce)
159 se <- sqrt(diag(info.r))
160
161 varyx0 <- 8.529411 # Var(Y/X=0) for standardized effect sizes; computed empirically
162
163 ## show effects, standard errors and effect sizes
164 round(cbind(est[23:43], se[23:43], est[23:43]/sqrt(varyx0)), 3)
165 round(cbind(ce[1:24], ce.se[1:24], ce[1:24]/sqrt(varyx0)), 3)

```

5 General Discussion

5.1 General Overview and Summary

In this thesis, we developed a new approach for estimating average and conditional treatment effects on count outcomes. The new moment-based approach aims at overcoming methodological limitations and shortcomings of earlier approaches (i.e., interpretation of regression coefficients, marginal effects) in three ways. First, the effect computation in the moment-based approach is based on the stochastic theory of causal effects (Steyer et al., 2014). Thus, treatment effects are unambiguously defined as the difference between (conditional) expectations under treatment and under control. If certain causality conditions hold, an effect can be interpreted as causal effect of the treatment. In contrast, interpreting exponentiated regression coefficients yields effects defined as ratios, which are more difficult to interpret from a causal perspective – especially if treatment-covariate interactions are involved. Second, the moment-based approach accounts for sampling variability in the observed covariates. In contrast to the marginal effects approach, where covariates are commonly assumed to be fixed-by-design (i.e., do not vary from sample to sample), we treat covariates as randomly sampled. This distinction yields an improved statistical inference and a gain in power, especially if strong interaction effects or non-normally distributed covariates are present. Third, covariates can be modeled as latent variables, which means that a covariate’s true scores are distinguished from measurement error in the indicator variables. The marginal effects approach only allows for observed values of a (fallible) covariate, which can – if measurement error is high – bias estimation of regression coefficients and, subsequently, average and conditional effects. In the moment-based approach, measurement error can be accounted for, which results in – in some cases also causally – unbiased estimates. In sum, the moment-based approach incorporates recent advancements in the statistical inference for treatment effects (see, e.g., Mayer et al., 2016) for count outcomes. In the following, we will summarize and discuss the central contributions of this thesis.

5.1.1 Treatment Effect Definition

In regressions with a logarithmic link function, the effectiveness of a treatment is commonly investigated using a ratio of conditional expectations. This effect measure is obtained by exponentiating estimated regression coefficients. In Chapter 1, we provided an excursus on definitions of ratio treatment effects from the perspective of the stochastic theory of causal effects. We discussed that the expectation (or “average”) of atomic ratio effects ($ARTE_{t0}$)

can only be estimated causally unbiased, if we have regressive independence of the outcome Y and the comprehensive potential confounder C_X , given values of the treatment variable X and an observed covariate Z . That means, randomization does not suffice as causality condition for the $ARTE_{t0}$, instead, for example, all covariates explaining unique variance in the outcome have to be included in order to get a causally unbiased effect estimate. The ratio of unconditional expectations of the true outcome variables RUC_{t0} – which can be represented by an exponentiated regression coefficient under some conditions – can be estimated causally unbiased, if a corresponding causality condition holds, but it is not an aggregate of atomic ratio treatment effects. Nevertheless, it is indicative for the average treatment effect ATE_{t0} and can therefore yield valuable information on the treatment's average effectiveness. In Chapter 2, we showed that in a simple example with a binary treatment variable and a single covariate, the $ARTE_{t0}$ and RUC_{t0} can be equivalent if some strong assumptions hold. That is, assuming the covariate Z takes the role of the comprehensive potential confounder C_X and does either have no effect on the outcome in the control group or does not alter the effect of the treatment (i.e., no interaction). In Chapter 3, we provided a graphical illustration of the incident rate ratio function of a single covariate Z . In regressions with a logarithmic link function and an interaction term, the incident rate ratio function is neither a constant nor does the ratio at the mean of the covariate equal the average ratio treatment effect $ARTE_{t0}$ or the RUC_{t0} . That is, mean-centering approaches – as commonly used in linear regression models (Cohen, Cohen, West, & Aiken, 2003) – do not necessarily yield causally interpretable quantities in regressions with a logarithmic link function.

One can find seemingly contradictory statements in the literature. For example, Gail et al. (1984) state that the ratio effects obtained by regression models with a logarithmic link function *are* collapsible. From our perspective, this contradiction can be resolved with a closer look at the underlying *true* regression model. Gail et al. (1984) do not include any treatment-covariate interactions in their regression model which relates to a generalization of Theorem 1 in Appendix 2.A from a single covariate to multiple covariates. Put differently – in the language of the stochastic theory of causal effects – the comprehensive potential confounder is assumed to either have no impact on the outcome in the control group or does not alter the effectiveness of the treatment. Both are very strong assumptions with regard to a comprehensive confounder, a variable that is literally defined as *all variables that influence choice or outcome of the treatment variable*. Thus, collapsibility of ratio effects only holds for special cases of regression models with a logarithmic link function, but not in general. For an overview on confounding and collapsibility, see Greenland et al. (1999).

5.1.2 Stochastic Covariates and Group Sizes

In the marginal effects approach, the covariates are usually assumed to be fixed-by-design, that is, their values are assumed to be predetermined by the experimenter and not ran-

domly sampled. While this might be a rare case in applied settings, adjustments of the standard errors are assumed to have only small effects on the corresponding inference on the marginal effects (e.g., [Wooldridge, 2010](#), p. 919). For average and conditional treatment effect analyses, a strong impact of stochastic covariates on statistical inference was found by [Kröhne \(2009\)](#). Standard errors derived under the fixed-covariate assumption were underestimated in conditions with non-zero interaction effects, resulting in inflated Type 1 error rates. Consequently, statistical procedures for estimating average and conditional treatment effects (e.g., [Li et al., 2019](#); [Mayer et al., 2016](#)) or other effect measures (e.g., simple slopes, [Y. Liu et al., 2017](#)) have adapted a stochastic covariates approach improving the statistical inference. In a similar vein, [Kröhne \(2009\)](#) pointed at the necessity to treat the treatment variable as random variable, unless the observed group sizes are actually predetermined by the experimenter. Recently, these findings were supported by [Mayer and Thoemmes \(2019\)](#), who investigated the issue of stochastic group sizes in factorial ANOVA designs. They could show that standard errors were underestimated when actually random group sizes were considered as predetermined. Thus, accounting for stochastic group sizes can improve statistical inferences as well.

In our simulation in Chapter 2, we also found the phenomenon of underestimated standard errors for marginal effects based on a regression model with a logarithmic link function. Type 1 error rates were inflated up to 40% in extreme conditions. The moment-based approach considers stochastic regressors and stochastic group weights, that is, both values of the covariates and the observed size of (treatment) groups can vary from sample to sample. In Chapter 2, we considered effect estimation using a single stochastic covariate. In Chapter 3, a multi-dimensional stochastic covariate was introduced by assuming a known multivariate distribution of its values, that is, a conditional multivariate normal distribution. We also extended the moment-based approach for estimating stochastic group weights using two different techniques. First, with a multinomial model the group probabilities are estimated with regard to a reference group and given the total sample size. Second, with a Poisson model the samples sizes are assumed to be Poisson distributed each. Thus, a difference between these two approaches is that the multinomial model considers the distribution of group sizes within a given total sample size, while the total sample size is random in the Poisson model. To our knowledge, there is no comprehensive comparison of the two approaches in the literature and, at least in our applied examples, we could not observe any significant differences between the estimated group probabilities and their corresponding standard errors.

5.1.3 Latent Covariates

In psychology and other social sciences, many constructs of interest are not directly observable, but latent variables. They are assessed through multiple indicator variables measuring the latent construct but also containing measurement error. It is well known that measurement error in covariates leads to attenuation bias in regression models ([Car-](#)

roll et al., 2010), decreasing the power to detect significant regressors or interaction terms. Moreover, measurement error can make a difference whether the causality condition of regressive independence for an unbiased effect function holds or not (Sengewald et al., 2018). Both effects of measurement error can be complex and unpredictable, when two or more latent variables are included in an analysis (see, e.g., Sengewald & Pohl, 2019). For an overview of the role of latent variables in causal effect analysis, see Mayer (2019). However, the marginal effects approach requires an empirical distribution of the covariates, which is not available if latent covariates are involved.

In contrast, the moment-based approach naturally incorporates latent covariates, because only few distributional parameters such as means and variances of latent variables are required for effect estimation. Using advanced structural equation modeling techniques, the simultaneous estimation of a count regression and a latent variable model is possible. The statistical framework, a negative binomial multigroup structural equation model, was presented in Chapter 3. We assumed (conditional) multivariate normally distributed covariates for both the model and the effect estimation. In an illustrative example, we found differential results depending on whether measurement error has been accounted for or not. When using a (fallible) mean score instead of a latent variable model, magnitude, direction, and significance of estimated regression coefficients changed. In this light, accounting for measurement error in covariates is a crucial feature of the moment-based approach.

5.1.4 Non-Normal Covariates

While the extension for stochastic covariates, stochastic group weights, and latent covariates are desirable for unbiased effect estimation and improvement of statistical inferences, they come at a price. In contrast to the marginal effects approach, which is based on the empirical distribution of the covariates, we had to assume (conditional multivariate) normally distributed covariate(s). In Chapter 2, we demonstrated that a violation of this assumption can lead to biased effect estimates and impaired inferences.

Thus, Chapter 4 was mainly concerned with a generalization of the moment-based approach to various kinds of covariates' joint distribution allowing for a more nuanced effect analysis. We proposed different ways to deal with non-normally distributed covariates, which can be categorized into two solutions. First, we distinguish between categorical and continuous covariates. For categorical covariates, a saturated model is fitted, which relieves us from an assumption on the distribution. This idea originates from Dietzfelbinger (2014) and is, for example, used in the EffectLiteR approach (Mayer et al., 2016). Second, for continuous covariates a joint distribution has to be specified, although not necessarily the multivariate normal distribution. Alternatives include skew-normal distributions, factorized joint distributions (e.g., marginal and conditional distribution), and approximations using finite Gaussian mixtures. In sum, these generalization make the moment-based approach applicable even when non-normally distributed covariates ought to be included.

5.1.5 Guidelines for Use of Moment-Based Approach

Despite their different statistical frameworks, the moment-based approach and the marginal effects approach need not necessarily lead to different results in applied settings. In our illustrative examples in Chapters 2 and 3 as well as in our Monte Carlo simulation study in Chapter 2, point estimates and statistical inference for average and conditional treatment effects were oftentimes very similar in both approaches. This finding might raise the question under which circumstances it should be (substantially) beneficial to use the moment-based approach and under which circumstances it may suffice to use the marginal effects approach. Thus, we want to provide some guidelines to answer this question.

First, a main goal of the moment-based approach is to provide improved statistical inference compared to the marginal effects approach. Consequently, considering stochastic group sizes, stochastic covariates, and latent covariates is first and foremost aimed at standard error estimation, but not necessarily point estimation. The moment-based approach accounts for various sources of uncertainty, which are due to the nature of the sampling process. If treatment groups and their size are generated randomly, if individuals and their covariate values are sampled randomly, and if the covariates are measured with error, all of this stochasticity should be incorporated into statistical inference. We recommend this procedure especially for high-stakes situations, where an accurate inference is crucial, for example, if a newly developed psychotherapy to reduce substance use disorders should be admitted by a statutory health insurance.

Second, in scenarios where only manifest covariates with no measurement error or high reliability are accounted for, the marginal effects approach could be a reasonable alternative to the moment-based approach. Given a sufficient sample size, the empirical distribution of the covariates yields a good approximation to the true distribution and, therefore, unbiased effect estimates. In contrast, the moment-based approach yields the risk of misspecifying the covariates distribution and, in turn, the risk of biased effect estimates. Thus, the risk of biased estimates has to be balanced against the benefits of stochastic covariates. However, in recent years, researchers proposed and discussed standard error formulas which allow for stochastic covariates in marginal effects, too (e.g., [Basu & Rathouz, 2005](#)). To our knowledge, marginal effects with stochastic covariates are not yet supported in any major statistical software, but may be a viable alternative in future research.

Third, there is one exception from the previous recommendation. If the covariates' true distribution is known or the risk of a misspecification of this distribution is at least very low, the moment-based approach performs substantially better in small samples both with regard to point and standard error estimation. In these scenarios, the empirical distribution which is used in the marginal effects approach does usually not sufficiently approximate the true distribution, potentially leading to biased point estimates and impaired inferences. In contrast, the moment-based approach incorporates a valuable information into the estimation procedure and thus yields better point estimates and standard errors.

Fourth, the statistical framework of the moment based approach (i.e., structural equation modeling) is more versatile than common generalized linear models, offering many features and modeling opportunities. Two of the ones we used in our illustrative examples are (a) flexible modeling of overdispersion in the negative binomial regression among multiple groups, which means the overdispersion parameter can vary among groups and (b) full information maximum likelihood estimation, which allows to deal with missing data without excluding cases. Further, models could be extended to include, for example, the estimation of latent classes as categorical covariates. In sum, if any of the more elaborate modeling techniques in structural equation modeling is required for effect analysis, the moment-based approach might be the method of choice.

5.2 Limitations and Future Research

The previous summary and guidelines point at several limitations of the moment-based approach, but also to possibilities of further development and future research. Along with the development and extension of the moment-based approach in Chapters 2, 3, and 4, we already discussed particular limitations and practical issues, for example, treatment non-compliance, handling of missing data, or multilevel structures. Thus, we will not repeat these remarks here, but want to discuss some more general aspects.

5.2.1 In-Depth Comparison of Effect Estimation Approaches

In the previous section, we provided some guidelines and comparison on the use of the moment-based and the marginal effects approach. These were based on the theoretical derivations and the simulation study included in this thesis and may offer a first overview and rough assessment of both approaches. However, a comprehensive Monte Carlo simulation study investigating the performances of both approaches will be indispensable to draw further conclusions. An extensive comparison should focus on four issues:

First, in applied data analyses usually a substantial number of covariates is included in a regression model, some of which are manifest, others are latent. While our theoretical derivations often address isolated phenomena (e.g., measurement error in covariates can lead to attenuation bias), these phenomena can interact in real-world settings. Thus, it is possible that in rather complex regression models, consequences of measurement error or the fixed-covariate assumption amplify each other – or compensate for each other. Similar effects have been reported by [Sengewald and Pohl \(2019\)](#), in models with two latent covariates. Accounting for the complexity of real-world data analyses and resulting amplification and compensation of stochastic regressors, measurement error, or violated distributional assumptions, might be crucial to derive valuable information on differential performance of the moment-based and the marginal effects approach in applied settings.

Second, a distinct investigation of experimental and quasi-experimental settings is advisable, as covariate selection and inclusion pursues different goals between these settings.

In experimental settings, unbiasedness of the conditional effect function is implied by randomization. For example, the average treatment effect can be unbiasedly estimated by a simple difference of group averages. Thus, the inclusion of covariates aims at increasing statistical power and efficiency. In the Chapters 3 and 4, we confined the illustration of the moment-based approach to randomized experiments. In quasi-experimental settings, unbiasedness of the conditional effect function can be achieved by covariate selection. Thus, the inclusion of covariates aims at obtaining unbiased effect estimates. Statistical power and efficiency can be considered second order qualities of interest in these settings. Accordingly, the moment-based and the marginal effects approach might perform distinctively good under different scenarios.

Third, a Monte Carlo investigation of standard error formulas for marginal effects with stochastic regressors is indicated. [Basu and Rathouz \(2005\)](#) proposed a formula for standard errors, which includes the variance of the estimated marginal effects in addition to the variance-covariance of the estimated regression coefficients. Later, this formula has been confirmed by [Terza \(2016\)](#) and [Wooldridge \(2010\)](#). Nevertheless, Wooldridge concludes that the formula may be technically correct, but “the adjustment may have a small effect” (p. 919). Therefore, it should be investigated whether the adjusted standard error for marginal effects yields equivalent statistical inference compared to the moment-based approach or not.

Fourth, one particular strength of the marginal effects approach is its parsimonious statistical framework, which is based on a generalized linear model with a logarithmic link function only. This framework should yield a relatively stable estimation of model parameters compared to the more elaborate NB-MG-SEM framework. For example, in Chapter 4, our NB-NG-SEM did not converge with group-specific overdispersion parameters and thus we had to fix these parameters to zero to obtain convergence. Numerical stability of the generalized linear model, even in presence of a multitude of covariates, could at least partly make up for other drawbacks of the marginal effects approach. Therefore, a Monte Carlo simulation should examine if the statistical inference for marginal effects is appropriate in scenarios where the moment-based approach or the NB-MG-SEM respectively exhibit convergence issues and numerical instability.

5.2.2 Non-Canonical Use of Link Functions

Throughout this thesis, we considered the logarithmic link function as the canonical link for count outcomes. However, it is possible to deviate from this statement in both directions: using a logarithmic link function for non-count outcomes and using other link functions for count outcomes.

First, the logarithmic link function can be and frequently is applied to investigate outcomes other than count variables. For real-valued outcomes with lower bound at zero, a generalized linear model with a logarithmic link function and Gamma distributed random component was embraced in recent years (e.g., [Ng & Cribbie, 2016](#); [Salway & Wakefield,](#)

2008), for example for reaction time data (Lo & Andrews, 2015). The moment-based approach can be directly adapted for such regression models, because it is basically build around the joint distribution of the covariates, not the outcome. That is, applying the moment-generating function works regardless of the conditional distribution of the outcome, whether Poisson, negative binomial, or Gamma.

Second, while the logarithmic link function is the canonical link for count outcomes, alternative link functions might be used under certain conditions. One alternative is the linear link function with Poisson or normally distributed random component, which can be useful if the expected counts are not approaching zero. Coxe et al. (2009) suggest expected counts larger than ten as a rule of thumb. From a statistical point of view, this approach is legitimate because the Poisson distribution $\mathcal{P}(\lambda)$ is well approximated by a normal distribution for large λ . If a linear link function is used, treatment effects can be obtained with the framework presented by Mayer et al. (2016).

5.2.3 Zero-Inflated Count Data and Hurdle Models

Sometimes, count data contain more zeros than a Poisson or negative binomial regression would reasonably predict. This can be due to so-called *structural zeros*, for example, if the frequency of smoking cigarettes in one day is examined and a substantial part of the sample consists of non-smokers. The observed count of zeros confounds the structural non-smokers and smokers who might not have smoked on that day. Lambert (1992) introduced a zero-inflated Poisson model, in which the conditional distribution of the outcome is characterized by a zero-component and a Poisson distribution. An alternative approach are hurdle models (Mullahy, 1986), which decompose the model into a logistic regression predicting zero vs. non-zero counts, and a truncated Poisson regression for the non-zero counts. For an expert tutorial on the analysis of zero-inflated count data, see Loeys, Moerkerke, Smet, and Buysse (2011).

In hurdle models, treatment effects could be computed with regard to two distinct models: a logistic regression and a truncated Poisson regression. Treatment effect estimation in logistic regression models is prone to similar issues that we treated for the logarithmic link function in this thesis. First, treatment effects are defined and estimated as odds ratios. These are difficult to interpret from a causal perspective (Hanmer & Kalkan, 2013) and non-collapsible, even in regression models without interaction term (Gail et al., 1984; Greenland et al., 1999). Second, treatment effects defined as differences can be obtained with a marginal effects approach. This approach, however, treats covariates and group sizes as fixed-by-design rather than randomly sampled. Consequences of the fixed-covariate assumption have been discussed at various points in this thesis. Third, logistic regression based on a generalized linear model framework does not allow to include latent variables and thus, measurement error in the covariates is neglected.

Consequently, it could be beneficial to adapt the extensions and advancements that we proposed for count regression models to logistic regression models. This would (not

only) allow to extend the moment-based approach for count outcomes to hurdle models and thereby have a better support for zero-inflated count data. Moreover, translating the moment-based approach to logistic regression models, yield applications wherever treatment effects on binary outcomes are investigated. An applied example would be examining the effectiveness of mathematical remedy courses to reduce higher education drop out among university freshmen (Tieben, 2019). However, there is no one-to-one adaption, because the moment-based approach depends on the logarithmic link function and the use of moment-generating functions, which cannot be used for the logistic link function.

5.2.4 Marginal Effects for Latent Covariates Revisited

Throughout this thesis, we often implied that marginal effects and latent variables are mutually incompatible. This seemingly incompatibility emerges from the contrast of the empirical distribution-based marginal effects to the generally non-observable true scores of a latent variable. However, there are ways in which – apart from factor scores – true scores of a latent variable could be obtained for marginal effect estimation.

Marginal maximum likelihood estimation for the NB-MG-SEM relies on the distributional parameters of the latent variables and marginalizes over their joint distribution. However, there are two alternative estimation frameworks for the NB-MG-SEM, which internally use estimated or sampled true scores to obtain parameter estimates. One is the EM algorithm (Dempster et al., 1977) computing the conditional expectation of the latent variables given values of all observed variables in each iteration. The other is Bayesian structural equation modeling (Muthén & Asparouhov, 2012), which samples values of the latent variables in each iteration. In Bayesian SEM it would be possible to apply marginal effect estimation within each iteration. That is, the sampled true scores and the observed covariates are used to compute conditional expectations under treatment and under control and, consequently, a marginal effect for each observation. Averaging over these marginal effects in each iteration of the MCMC estimation results in a posterior distribution of the average treatment effect. The expectation of this posterior distribution can usually be interpreted as estimated average treatment effect, its standard deviation as corresponding standard error.

However, using Bayesian structural equation models for effect estimation may not be straightforward in all cases. First, MacCallum, Edwards, and Cai (2012) refer to various computational issues of the MCMC algorithm in latent variable models. For example, the posterior distribution (e.g., of the average treatment effect) might have multiple modes making a careful investigation of the posterior distribution necessary. Further, the identification of the model can be difficult to determine, because the states of identification and non-identification seamlessly blend into each other at some point. In addition, the identification as well as the convergence of the estimated parameters have to be carefully investigated, which can be difficult for inexperienced users. Second, the specification of

prior distribution for each parameter can have an impact on the resulting parameter estimates. In a simulation study, [Van Erp, Mulder, and Oberski \(2018\)](#) found that three different default priors can lead to different results demanding a careful priors selection. For an overview of further cautions (and hopes) in Bayesian SEM, see [MacCallum et al. \(2012\)](#).

5.2.5 Treatment Effects Defined as Ratios

We provided an excursus and several small theorems on the definition and properties of treatment effects defined as ratios. Our findings indicate that the properties of ratio effects strongly differ from the properties of difference effects, for example, with regard to expectation stability and collapsibility. However, a comprehensive treatment of this topic from a causal inference perspective would be desirable. Analog to the theoretical derivations of [Steyer et al. \(2014\)](#) and [Mayer, Thoemmes, et al. \(2014\)](#), a mathematical definition of atomic ratio treatment effects and derivation of their aggregation to average and conditional effects, their decomposition into total, direct, and indirect effects, and causality conditions implying unbiasedness of the corresponding ratio effect function would foster a unified overview of issues related to ratio effects.

To date, the complexities of ratio effects have largely been ignored in the causal effects literature. For example, [Rubin \(2005\)](#) confines himself to the notion that “the ratio or any such comparison could be used to define unit-level causal effects.” (p. 323), before continuing investigation of difference effects. [Greenland \(1996\)](#) addresses the issue of collapsibility in ratio effects. In epidemiology, the difficulties to define a *true* ratio effect given the several possibilities to specify causal and non-causal ratio effects, have been presented by [Maldonado and Greenland \(2002\)](#). However, to our knowledge, a comprehensive investigation of aggregation of atomic to conditional effects and the question of causality conditions implying unbiasedness has not yet been addressed in the literature.

5.3 Concluding Remarks

The main goal of this thesis was to incorporate recent advancements in the statistical inference on treatment effects (see, e.g., [Mayer et al., 2016](#)) for count outcomes. The development of the moment-based approach gives researchers the opportunity to apply complex and nuanced treatment effect analysis on their count outcomes using modern techniques of statistical modeling and inference. Many of the features described in this thesis have already been condensed into an open-source R package named *CountEffects* ([Kiefer & Mayer, 2020](#)). Further features are about to be added in order to make the moment-based approach accessible for a broad range of applied settings.

References

- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
- Ahern, J., Hubbard, A., & Galea, S. (2009). Estimating the effects of potential public health interventions on population disease burden: A step-by-step illustration of causal inference methods. *American Journal of Epidemiology*, 169(9), 1140–1147. doi: 10.1093/aje/kwp015
- Aiken, L. S., Mistler, S. A., Cox, S., & West, S. G. (2015). Analyzing count variable in individual and groups: Single level and multilevel models. *Group Processes and Intergroup Relations*(3), 290–314. doi: 10.1177/1368430214556702
- Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422), 669–679. doi: 10.1080/01621459.1993.10476321
- Albert, J. M., & Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, 67(3), 1028–1038. doi: 10.1111/j.1541-0420.2010.01547.x
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444–455. doi: 10.1080/01621459.1996.10476902
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum.
- Austin, P. C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics*, 5(1). doi: 10.2202/1557-4679.1146
- Austin, P. C., Stryhn, H., Leckie, G., & Merlo, J. (2017). Measures of clustering and heterogeneity in multilevel poisson regression analyses of rates/count data. *Statistics in Medicine*, 37(4), 572–589. doi: 10.1002/sim.7532
- Azzalini, A., & Capitanio, A. (2009). *The skew-normal and related families*. Cambridge University Press. doi: 10.1017/cbo9781139248891
- Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4), 715–726. doi: 10.1093/biomet/83.4.715
- Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., . . . Willis, S. L. (2002). Effects of cognitive training interventions with older adults. *JAMA*, 288(18), 2271. doi: <https://doi.org/10.1001/jama.288.18.2271>
- Barnard, J., Frangakis, C. E., Hill, J. L., & Rubin, D. B. (2003). Principal stratification

- approach to broken randomized experiments. *Journal of the American Statistical Association*, 98(462), 299–323. doi: 10.1198/016214503000071
- Basu, A., & Rathouz, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics*, 6(1), 93–109. doi: 10.1093/biostatistics/kxh020
- Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology*, 31, 419–456. doi: 10.1146/annurev.ps.31.020180.002223
- Blanca, M. J., Arnau, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, 9(2), 78–84. doi: 10.1027/1614-2241/a000057
- Bollen, K. A. (1989). *Structural equation modeling with latent variables*. New York, NY: Wiley.
- Bono, R., Blanca, M. J., Arnau, J., & Gómez-Benito, J. (2017, sep). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, 8. doi: 10.3389/fpsyg.2017.01602
- Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference - Theory and methods*. New York, USA: Springer.
- Carota, C. (2002). Semiparametric regression for count data. *Biometrika*, 89(2), 265–281. doi: 10.1093/biomet/89.2.265
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2010). *Measurement error in nonlinear models: A modern perspective*. Boca Raton, FL: Chapman and Hall/CRC.
- Cascione, M., Tenenbaum, N., Wendt, J., Meng, X., Schofield, L., & Cree, B. A. (2018). Treatment retention on fingolimod compared with injectable multiple sclerosis therapies in African-American patients: A subgroup analysis of a randomized phase 4 study. *Multiple Sclerosis and Related Disorders*, 25, 50–56. doi: 10.1016/j.msard.2018.07.014
- Chen, X. (2006). The adjustment of random baseline measurements in treatment effect estimation. *Journal of Statistical Planning and Inference*, 136, 4161–4175. doi: 10.1016/j.jspi.2005.08.046
- Cheng, J., Cheng, N. F., Guo, Z., Gregorich, S., Ismail, A. I., & Gansky, S. A. (2017). Mediation analysis for count and zero-inflated count data. *Statistical Methods in Medical Research*, 27(9), 2756–2774. doi: 10.1177/0962280216686131
- 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.
- Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330–351. doi: 10.1037//1082-989x.6.4.330-351
- Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to poisson regression and its alternatives. *Journal of Personality Assessment*,

- 91(2), 121-136. doi: 10.1080/00223890802634175
- Crager, M. R. (1987). Analysis of covariance in parallel-group clinical trials with pretreatment baselines. *Biometrics*, 43, 895-901. doi: 10.2307/2531543
- Cree, B. A., Arnold, D. L., Cascione, M., Fox, E. J., Williams, I. M., Meng, X., ... Tenenbaum, N. (2018). Phase IV study of retention on fingolimod versus injectable multiple sclerosis therapies: a randomized clinical trial. *Therapeutic Advances in Neurological Disorders*, 11, 175628641877433. doi: 10.1177/1756286418774338
- Daniel, R. M., Stavola, B. L. D., & Cousens, S. N. (2011). Gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *The Stata Journal: Promoting communications on statistics and Stata*, 11(4), 479-517. doi: 10.1177/1536867x1201100401
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 39(1), 1-38. doi: 10.1111/j.2517-6161.1977.tb01600.x
- Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, 9. doi: 10.3389/fpsyg.2018.00580
- DeSantis, S. M., Lazaridis, C., Ji, S., & Spinale, F. G. (2013). Analyzing propensity matched zero-inflated count outcomes in observational studies. *Journal of Applied Statistics*, 41(1), 127-141. doi: 10.1080/02664763.2013.834296
- Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, 76(5), 741-770. doi: 10.1177/0013164415607618
- Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*, 13, 31-38. doi: 10.1027/1614-2241/a000130
- Dietzfelbinger, L. (2014). *Generalisierte Kovarianzanalyse mit qualitativen und metrischen Kovariaten zur Analyse durchschnittlicher und bedingter Effekte [Generalized analysis of covariance with qualitative and metric covariates for the analysis of average and conditional effects]*. Jena, Germany: Unpublished diploma thesis, Friedrich-Schiller-Universität Jena, Germany.
- Dobson, A. (2001). *An introduction to generalized linear models*. Boca Raton, FL: Chapman & Hall.
- Dowd, B. E., Greene, W. H., & Norton, E. C. (2014). Computation of standard errors. *Health Services Research*, 49(2), 731-750. doi: 10.1111/1475-6773.12122
- Dunson, D. B. (2005). Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470), 618-627. doi: 10.1198/016214504000001457
- El-Sayyad, G. M. (1973). Bayesian and classical analysis of Poisson regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(3), 445-451. doi: 10.1111/j.2517-6161.1973.tb00972.x
- Enders, C. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

- Ertefaie, A., & Stephens, D. A. (2010). Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics*, 6(2). doi: 10.2202/1557-4679.1198
- Gail, M. H., Wieand, S., & Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71, 431–444. doi: 10.1093/biomet/71.3.431
- Gardner, W., Mulvey, E. P., & Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3), 392–404. doi: 10.1037//0033-2909.118.3.392
- Garrett, S. B., Doyle, S. R., Peavy, K. M., Wells, E. A., Owens, M. D., Shores-Wilson, K., ... Donovan, D. M. (2018). Age differences in outcomes among patients in the “stimulant abuser groups to engage in 12-step” (STAGE-12) intervention. *Journal of Substance Abuse Treatment*, 84, 21–29. doi: 10.1016/j.jsat.2017.10.012
- Gatsonis, C., & Sampson, A. R. (1989). Multiple correlation: Exact power and sample size calculations. *Psychological Bulletin*, 106, 516–524. doi: 10.1037/0033-2909.106.3.516
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). London, England: Chapman & Hall.
- Geneletti, S., & Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 728–749). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199574131.003.0034
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3–8. doi: 10.2307/1174772
- Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18, 36–56. doi: 10.1093/pan/mpp036
- Graham, J. W. (2012). *Missing data: Analysis and design*. New York, NY: Springer.
- Greene, W. (2007). *Econometric analysis*. Upper Saddle River, NJ: Prentice-Hall.
- Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, 7(5), 498–501. doi: 10.1097/00001648-199609000-00008
- Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14, 29–46. doi: 10.1214/ss/1009211805
- Guo, Z., Small, D. S., Gansky, S. A., & Cheng, J. (2017). Mediation analysis for count and zero-inflated count data without sequential ignorability and its application in dental studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67(2), 371–394. doi: 10.1111/rssc.12233
- Gupta, A. K., Gonzalez-Farias, G., & Dominguez-Molina, J. (2004). A multivariate skew normal distribution. *Journal of Multivariate Analysis*, 89(1), 181–190. doi: 10.1016/s0047-259x(03)00131-3
- Gurmu, S., Rilstone, P., & Stern, S. (1999). Semiparametric estimation of count regression

- models. *Journal of Econometrics*, 88(1), 123–150. doi: 10.1016/s0304-4076(98)00026-8
- Hanmer, M. J., & Kalkan, K. O. (2013). Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science*, 57(1), 263–277.
- Helseth, S. A., Waschbusch, D. A., Gnagy, E. M., Onyango, A. N., Burrows-MacLean, L., Fabiano, G. A., . . . Pelham, W. E. (2015). Effects of behavioral and pharmacological therapies on peer reinforcement of deviancy in children with adhd-only, adhd and conduct problems, and controls. *Journal of Consulting and Clinical Psychology*, 83(2), 280–292. doi: 10.1037/a0038505
- Hilbe, J. M. (2007). *Negative binomial regression*. New York, NY: Cambridge.
- Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). New York, NY: Cambridge.
- Hirano, K., & Imbens, G. W. (2002). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology*, 2, 259–278.
- Hittner, J. B., Owens, E. C., & Swickert, R. J. (2016). Influence of social settings on risky sexual behavior. *SAGE Open*, 6(1). doi: 10.1177/2158244016629187
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8). doi: 10.18637/jss.v042.i08
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960. doi: 10.1080/01621459.1986.10478354
- Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, 319(7211), 670–674. doi: 10.1136/bmj.319.7211.670
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 73–101. doi: 10.1007/978-1-4612-4380-9_35
- Huber, P. J., & Ronchetti, E. M. (2009). *Robust statistics*. New York, NY: Wiley.
- Imai, K., King, G., & Lau, O. (2008). Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics*, 17(4), 892–913. doi: 10.1198/106186008x384898
- Imbens, G. W. (2014). Instrumental variables: An econometrician’s perspective. *Statistical Science*, 29(3), 323–358. doi: 10.1214/14-sts480
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139025751
- Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling*, 8(2), 205–223. doi: 10.1207/s15328007sem08023
- Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, 10 (1), 128–141. doi: 10.1207/s15328007sem1001_6

- Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30, 175–193. doi: 10.1007/bf02418571
- Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., ... Kleinman, K. (2001). ACTIVE: A cognitive intervention trial to promote independence in older adults. *Controlled Clinical Trials*, 22(4), 453–479. doi: 10.1016/S0197-2456(01)00139-8
- Johns, G., & Al Hajj, R. (2016). Frequency versus time lost measures of absenteeism: Is the voluntariness distinction an urban legend. *Journal of Organizational Behavior*, 37(3), 456–479. doi: 10.1002/job.2055
- Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, 1, 57–93.
- Kang, J. D. Y., & Schafer, J. L. (2006). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean for incomplete data. *Statistical Science*, 22, 523–539. doi: 10.1214/07-sts227
- Karrison, T., & Kocherginsky, M. (2018, mar). Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials? *Clinical Trials*, 15(2), 178–188. doi: 10.1177/1740774518759281
- Kelava, A., & Brandt, H. (2014). A general nonlinear multilevel structural equation mixture model. *Frontiers in Psychology*, 5(748). doi: <http://dx.doi.org/10.3389/fpsyg.2014.00748>
- Kelava, A., & Nagengast, B. (2012). A Bayesian model for the estimation of latent interaction and quadratic effects when latent variables are non-normally distributed. *Multivariate Behavioral Research*, 47(5), 717–742. doi: 10.1080/00273171.2012.715560
- Kelava, A., Nagengast, B., & Brandt, H. (2014). A nonlinear structural equation mixture modeling approach for nonnormally distributed latent predictor variables. *Structural Equation Modeling*, 21, 468–481. doi: 10.1080/10705511.2014.915379
- Kenny, T. E., Van Wijk, M., Singleton, C., & Carter, J. C. (2018). An examination of the relationship between binge eating disorder and insomnia symptoms. *European Eating Disorders Review*, 26(3), 186–196. doi: 10.1002/erv.2587
- Kiefer, C., & Mayer, A. (2019). Average effects based on regressions with a logarithmic link function: A new approach with stochastic covariates. *Psychometrika*, 84(2), 422–446. doi: 10.1007/s11336-018-09654-1
- Kiefer, C., & Mayer, A. (2020). *CountEffects: An R package to computed average and conditional effects for count outcomes*. Retrieved March 31, 2020, from <https://github.com/chkiefer/countEffects>.
- Kiefer, C., & Mayer, A. (in press). Accounting for latent covariates in average effects from count regressions. *Multivariate Behavioral Research*. doi: 10.1080/00273171.2020.1751027
- Kiefer, C., Rosseel, Y., Wiese, B. S., & Mayer, A. (2018). Modeling and predicting non-linear changes in educational trajectories: The multilevel latent growth components approach. *Psychological Test and Assessment Modeling*, 60(2), 189–221.

- King, G., & Zeng, L. (2006). The dangers of extreme counterfactuals. *Political Analysis*, 14, 131–159. doi: 10.1093/pan/mpj004
- Kröhne, U. (2009). *Estimation of average causal effects in quasi-experimental designs: Non-linear constraints in structural equation models* (Unpublished doctoral dissertation). Friedrich-Schiller-University Jena, Germany.
- Kukush, A., Schneeweis, H., & Wolf, R. (2004). Three estimators for the Poisson regression model with measurement errors. *Statistical Papers*, 45(3), 351–368. doi: 10.1007/bf02777577
- Laborde-Castérot, H., Agrinier, N., & Thilly, N. (2015). Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: A systematic review. *Journal of Clinical Epidemiology*, 68(10), 1232–1240. doi: 10.1016/j.jclinepi.2015.04.003
- Lachman, M. E., Baltes, P., Nesselroade, J. R., & Willis, S. L. (1982). Examination of personality-ability relationships in the elderly: The role of the contextual (interface) assessment mode. *Journal of Research in Personality*, 16(4), 485–501. doi: 10.1016/0092-6566(82)90007-1
- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1). doi: 10.2307/1269547
- Layland, E. K., Calhoun, B. H., Russell, M. A., & Maggs, J. L. (2018). Is alcohol and other substance use reduced when college students attend alcohol-free programs? Evidence from a measurement burst design before and after legal drinking age. *Prevention Science*, 1–11. doi: 10.1007/s11121-018-0877-6
- Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, England: Wiley.
- Leeper, T. J. (2018). margins: Marginal effects for model objects [Computer software manual]. (R package version 0.3.25)
- Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, 12, 1–27. doi: 10.1207/s15328007sem1201_1
- 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>.
- Li, L. (2010). Comment: Analyzing propensity score matched count data. *The International Journal of Biostatistics*, 6(1). doi: 10.2202/1557-4679.1214
- Li, L., McLouth, C. J., & Delaney, H. D. (2019). Analysis of covariance in randomized experiments with heterogeneity of regression and a random covariate: The variance of the estimated treatment effect at selected covariate values. *Multivariate Behavioral Research*, 1–15. doi: 10.1080/00273171.2019.1693953
- Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique. *The Annals of Applied Statistics*, 7(1), 295–318. doi: 10.1214/12-aos583

- Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken, NJ: Wiley.
- Liu, C. (2005). Robit regression: A simple robust alternative to logistic and probit regression. In A. Gelman & X.-L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 227–238). Hoboken, NJ: John Wiley & Sons, Ltd. doi: 10.1002/0470090456.ch21
- Liu, Y., & West, S. G. (2015). Weekly cycles in daily report data: An overlooked issue. *Journal of Personality*, 84(5), 560–579. doi: 10.1111/jopy.12182
- Liu, Y., West, S. G., Levy, R., & Aiken, L. S. (2017). Tests of simple slopes in multiple regression models with an interaction: Comparison of four approaches. *Multivariate Behavioral Research*, 52(4), 1–20. doi: 10.1080/00273171.2017.1309261
- Lo, S., & Andrews, A. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6, 1171. doi: 10.3389/fpsyg.2015.01171
- Loeys, T., Moerkerke, B., Smet, O. D., & Buysse, A. (2011). The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression. *British Journal of Mathematical and Statistical Psychology*, 65(1), 163–180. doi: 10.1111/j.2044-8317.2011.02031.x
- Lohman, M. C., Rebok, G. W., Spira, A. P., Parisi, J. M., Gross, A. L., & Kueider, A. M. (2012). Depressive symptoms and memory performance among older adults. *Journal of Aging and Health*, 25, 209–229. doi: 10.1177/0898264312460573
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables. advanced quantitative techniques in the social sciences*. Thousand Oaks, CA: Sage Publications.
- Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, 57, 581–597. doi: 10.1007/bf02294421
- Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, 38(4), 751–766. doi: 10.1016/j.aap.2006.02.001
- Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, 15(1). doi: 10.1186/s12982-018-0069-7
- MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and cautions in implementing Bayesian structural equation modeling. *Psychological Methods*, 17(3), 340–345. doi: 10.1037/a0027131
- Macdonald, K., Germine, L., Anderson, A., Christodoulou, J., & McGrath, L. M. (2017). Dispelling the myth: Training in education or neuroscience decreases but does not eliminate beliefs in neuromyths. *Frontiers in Psychology*, 8, 1314. doi: 10.3389/fpsyg.2017.01314
- Maldonado, G., & Greenland, S. (2002). Estimating causal effects. *International Journal of Epidemiology*, 31, 422–429. doi: 10.1093/ije/31.2.422

- Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013, sep). Why item parcels are (almost) never appropriate: Two wrongs do not make a right –camouflaging misspecification with item parcels in CFA models. *Psychological Methods*, 18(3), 257–284. doi: 10.1037/a0032773
- Mayer, A. (2019). Causal effects based on latent variable models. *Methodology*, 15, 15–28. doi: 10.1027/1614-2241/a000174
- Mayer, A., Dietzfelbinger, L., Rosseel, Y., & Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behavioral Research*, 51, 374–391. doi: 10.1080/00273171.2016.1151334
- Mayer, A., Nagengast, B., Fletcher, J., & Steyer, R. (2014). Analyzing average and conditional effects with multigroup multilevel structural equation models. *Frontiers in Psychology*, 5. doi: 10.3389/fpsyg.2014.00304
- Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. *Structural Equation Modeling*, 19, 513–533. doi: 10.1080/10705511.2012.713242
- Mayer, A., & Thoemmes, F. (2019). Analysis of variance models with stochastic group weights. *Multivariate Behavioral Research*, 54(4), 542–554. doi: 10.1080/00273171.2018.1548960
- Mayer, A., Thoemmes, F., Rose, N., Steyer, R., & West, S. G. (2014). Theory and analysis of total, direct and indirect causal effects. *Multivariate Behavioral Research*, 49(5), 425–442. doi: 10.1080/00273171.2014.931797
- Mayer, A., Umbach, N., Flunger, B., & Kelava, A. (2017). Effect analysis using nonlinear structural equation mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(4), 556–570. doi: 10.1080/10705511.2016.1273780
- Mazerolle, L., Bennett, S., Antrobus, E., Cardwell, S. M., Eggins, E., & Piquero, A. R. (2018). Disrupting the pathway from truancy to delinquency: A randomized field trial test of the longitudinal impact of a school engagement program. *Journal of Quantitative Criminology*. doi: 10.1007/s10940-018-9395-8
- McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, 58, 110–133. doi: 10.2307/1130295
- McCullagh, P., & Nelder, J. A. (1996). *Generalized Linear Models* (2nd ed.). Cambridge: Chapman & Hall.
- McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: John Wiley & Sons. doi: 10.1002/0471721182
- McNeish, D. (2019). Poisson multilevel models with small samples. *Multivariate Behavioral Research*, 54(3), 444–455. doi: 10.1080/00273171.2018.1545630
- Meier, P. (1991). Compliance as an explanatory variable in clinical trials: Comment. *Journal of the American Statistical Association*, 86(413), 19. doi: 10.2307/2289709
- Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85(4). doi: 10.18637/jss.v085.i04

- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156–166. doi: 10.1037/0033-2909.105.1.156
- Miratrix, L. W., Sekhon, J. S., & Yu, B. (2012). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2), 369–396. doi: 10.1111/j.1467-9868.2012.01048.x
- Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of missing data methodology*. New York, NY: CRC Press.
- Muench, F., van Stolk-Cooke, K., Kuerbis, A., Stadler, G., Baumel, A., Shao, S., ... Morgenstern, J. (2017). A randomized controlled pilot trial of different mobile messaging interventions for problem drinking compared to weekly drink tracking. *PLOS ONE*, 12(2), e0167900. doi: 10.1371/journal.pone.0167900
- Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33(3), 341–365. doi: 10.1016/0304-4076(86)90002-3
- Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics*, 79(4), 586–593. doi: 10.1162/003465397557169
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313–335. doi: 10.1037/a0026802
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376–398. doi: 10.1177/0049124194022003006
- Muthén, L. K., & Muthén, B. O. (1998-2015). *Mplus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- 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.
- Neighbors, C., Rodriguez, L. M., Rinker, D. V., Gonzales, R. G., Agana, M., Tackett, J. L., & Foster, D. W. (2015). Efficacy of personalized normative feedback as a brief intervention for college student gambling: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, 83(3), 500–511. doi: 10.1037/a0039125
- Ng, V. K. Y., & Cribbie, R. A. (2016). Using the gamma generalized linear model for modeling continuous, skewed and heteroscedastic outcomes in psychology. *Current Psychology*, 36(2), 1-11. doi: 10.1007/s12144-015-9404-0
- Nusser, L., & Weinert, S. (2017). Appropriate test-taking instructions for students with special educational needs. *Journal of Cognitive Education and Psychology*, 16(3), 227–240. doi: 10.1891/1945-8959.16.3.227
- O'Rourke, H. P., & Vazquez, E. (2019). Mediation analysis with zero-inflated substance use outcomes: Challenges and recommendations. *Addictive Behaviors*, 94, 16–25. doi: 10.1016/j.addbeh.2019.01.034
- Parisi, J. M., Franchetti, M. K., Rebok, G. W., Spira, A. P., Carlson, M. C., Willis, S. L.,

- & Gross, A. L. (2014). Depressive symptoms and inductive reasoning performance: Findings from the ACTIVE reasoning training intervention. *Psychology and Aging, 29*(4), 843–851. doi: 10.1037/a0037670
- Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge, UK. doi: 10.1017/cbo9780511803161
- Pelham, Jr., W. E., Fabiano, G. A., Waxmonsky, J. G., Greiner, A. R., Gnagy, E. M., Pelham, III, W. E., . . . Murphy, S. A. (2016). Treatment sequencing for childhood ADHD: a multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology, 45*(4), 396–415. doi: 10.1080/15374416.2015.1105138
- Precht, L., Keinath, A., & Krems, J. F. (2017). Effects of driving anger on driver behavior—results from naturalistic driving data. *Transportation Research Part F: Traffic Psychology and Behaviour, 45*, 75–92. doi: 10.1016/j.trf.2016.10.019
- Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics, 38*(2), 485. doi: 10.2307/2530463
- R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*, 167–190. doi: 10.1007/bf02295939
- Radloff, L. S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401. doi: 10.1177/014662167700100306
- Raykov, T. (2012). Propensity score analysis with fallible covariates: A note on a latent variable modeling approach. *Educational and Psychological Measurement, 72*, 715–733. doi: 10.1177/0013164412440999
- 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. doi: 10.1207/s15328007sem1104\7
- Reichhardt, C. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford: New York, NY.
- Rencher, A. C., & Schaalje, G. B. (2007). *Linear models in statistics* (2nd ed.). New York, NY: Wiley.
- Robins, J. (1986). A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modelling, 7*, 1393–1512.
- 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. doi: 10.1080/01621459.1994.10476818
- Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association, 90*, 106–121. doi: 10.1080/01621459.1995.10476493

- Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique*, 59(2), 227. doi: 10.2307/1403444
- Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.
- Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, 102, 191-200. doi: 10.1198/016214506000001112
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. doi: 10.18637/jss.v048.i02
- Rosset, S., & Tibshirani, R. J. (2018). From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115(529), 138-151. doi: 10.1080/01621459.2018.1424632
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701. doi: 10.1037/h0037350
- Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 322-331. doi: 10.1198/016214504000001880
- Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52, 249-264. doi: 10.2307/2533160
- Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: Statistical approaches and design issues. *Psychological Methods*, 19(3), 317-333. doi: 10.1037/met0000013
- Salway, R., & Wakefield, J. (2008). Gamma generalized linear models for pharmacokinetic data. *Biometrics*, 64(2), 620-626. doi: 10.1111/j.1541-0420.2007.00897.x
- Sampson, A. R. (1974). A tale of two regressions. *Journal of the American Statistical Association*, 69, 682-689. doi: 10.2307/2286002
- Schafer, J. L., & Kang, J. D. Y. (2008). Average causal effects from observational studies: A practical guide and simulated example. *Psychological Methods*, 13(4), 279-313. doi: 10.1037/a0014268
- Schaumborg, R. L., & Flynn, F. J. (2017). Clarifying the link between job satisfaction and absenteeism: The role of guilt proneness. *Journal of Applied Psychology*, 102(6), 982. doi: 10.1037/apl0000208
- Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8(1), 341-377. doi: 10.1146/annurev-economics-080315-015058
- Schrimshaw, E. W., Antebi-Gruszka, N., & Downing, J., M J. (2016). Viewing of internet-based sexually explicit media as a risk factor for condomless anal sex among men who have sex with men in four us cities. *PLoS one*, 11(4). doi: 10.1371/journal.pone

- .0154439
- Schuler, M. S., & Rose, S. (2016). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1), 65–73. doi: 10.1093/aje/kww165
- Sengewald, M.-A., & Pohl, S. (2019). Compensation and amplification of attenuation bias in causal effect estimates. *Psychometrika*, 84(2), 589–610. doi: 10.1007/s11336-019-09665-6
- Sengewald, M.-A., Steiner, P. M., & Pohl, S. (2018). When does measurement error in covariates impact causal effect estimates? analytic derivations of different scenarios and an empirical illustration. *British Journal of Mathematical and Statistical Psychology*, 72(2), 244–270. doi: 10.1111/bmsp.12146
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sheiner, L. B., & Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clinical Pharmacology & Therapeutics*, 57, 6–15. doi: 10.1016/0009-9236(95)90260-0
- Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7), 731–738. doi: 10.1093/aje/kwq472
- Splawa-Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (Reprinted from *roczniki nauk rolniczych* tom 10, pp. 1–51, 1923). *Statistical Science*, 5, 465–480.
- Sridharan, V., Shoda, Y., Heffner, J., & Bricker, J. (2019). A pilot randomized controlled trial of a web-based growth mindset intervention to enhance the effectiveness of a smartphone app for smoking cessation. *JMIR mHealth and uHealth*, 7(7), e14602. doi: 10.2196/14602
- Stefanski, L. A., & Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56(1), 29–38. doi: 10.1198/000313002753631330
- Sterba, S. K., & Rights, J. D. (2016). Accounting for parcel-allocation variability in practice: Combining sources of uncertainty and choosing the number of allocations. *Multivariate Behavioral Research*, 51(2-3), 296–313. doi: 10.1080/00273171.2016.1144502
- Steyer, R. (2005). Analyzing individual and average causal effects via structural equation models. *Methodology*, 1, 39–54. doi: 10.1027/1614-1881.1.1.39
- Steyer, R., Gabler, S., von Davier, A. A., & Nachtigall, C. (2000). Causal regression models II: Unconfoundedness and causal unbiasedness. *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., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In Michalos, A. C. (Ed.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 606–631). Dordrecht, Netherlands: Springer.
- Steyer, R., & Nagel, W. (2017). *Probability and conditional expectation*. Somerset, NJ : John Wiley & Sons, Incorporated.
- Stoll, H., King, G., & Zeng, L. (2006). WhatIf: RSoftware for evaluating counterfactuals. *Journal of Statistical Software*, 15(4). doi: 10.18637/jss.v015.i04
- Terza, J. V. (2016). Inference using sample means of parametric nonlinear data transformations. *Health Services Research*, 51(3), 1109–1113. doi: 10.1111/1475-6773.12494
- Tieben, N. (2019). Brückenkursteilnahme und Studienabbruch in ingenieurwissenschaftlichen Studiengängen. *Zeitschrift für Erziehungswissenschaft*, 22(5), 1175–1202. doi: 10.1007/s11618-019-00906-z
- Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. New York, NY: Springer.
- van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2. doi: 10.2202/1557-4679.1043
- Van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. *Psychological Methods*, 23(2), 363–388. doi: 10.1037/met0000162
- Wang, A., Nianogo, R. A., & Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology*, 17(1). doi: 10.1186/s12874-016-0282-4
- West, S. G., Cham, H., & Liu, Y. (2014). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & M. Judd (Eds.), *Handbook of research methods in social and personality psychology*. New York, NY: Cambridge.
- Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12, 308–331.
- Wolinsky, F. D., Weg, M. W. V., Martin, R., Unverzagt, F. W., Willis, S. L., Marsiske, M., ... Tennstedt, S. L. (2009). Does cognitive training improve internal locus of control among older adults? *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 65B(5), 591–598. doi: 10.1093/geronb/gbp117
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT.
- Zahodne, L. B., Meyer, O. L., Choi, E., Thomas, M. L., Willis, S. L., Marsiske, M., ... Parisi, J. M. (2015). External locus of control contributes to racial disparities in memory and reasoning training gains in ACTIVE. *Psychology and Aging*, 30(3), 561–572. doi: 10.1037/pag0000042