Check for updates

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2021)
© *The Authors. British Journal of Mathematical and Statistical Psychology*
published by John Wiley & Sons Ltd
on behalf of British Psychological Society

www.wileyonlinelibrary.com

Treatment effects on count outcomes with non-normal covariates

Christoph Kiefer* and Axel Mayer
Bielefeld University, Bielefeld, Germany

The effects of a treatment or an intervention on a count outcome are 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: an alternative, known non-normal distribution; a plausible factorization of the joint distribution; and 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.

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, Owens, & Swickert, 2016; Jobe et al., 2001; Mazerolle et al., 2019; Nusser & Weinert, 2017; Schaumberg & 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 and 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).

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

*Correspondence should be addressed to Christoph Kiefer, Bielefeld University, Bielefeld, Germany (email: christoph.kiefer@uni-bielefeld.de).
The data used in the illustrative example are publicly available in OpenICPSR at https://www.openicpsr.org/openicpsr/project/128941; https://doi.org/10.3886/E128941

DOI:10.1111/bmsp.12237

Recently, a moment-based approach was proposed by Kiefer and Mayer (2019, 2020) enabling aggregated effects to be computed using parameters of the (joint) distribution of the covariates (e.g., means, variances, and covariances). Instead of computing conditional effects given the observed covariate values and averaging over their empirical distribution, the conditional effects are integrated over all possible values of the covariates weighted by their likelihood given by the (assumed) joint density. While this procedure would usually have to be carried out by numerically solving an improper integral (e.g., an integral over conditional effects weighted by a multivariate normal density), the improper integral is replaced with a moment-generating function in the moment-based approach. Thus, as most moment-generating functions have closed-form solutions, the moment-based approach allows for a fast and analytical computation of the aggregated treatment effects.

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. In contrast, the traditional approach treats the observed covariate values as fixed by design (i.e., not varying between samples), which can lead to an underestimation of standard errors for the aggregated 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 (2020) 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 data sets 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 missspecification 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 (or metric) covariates (e.g., pre-test count variables). 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 finer-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 modelled and examined for more than one treatment condition (in comparison to a control group).

This paper is structured as follows. First, we provide general definitions of treatment effects on count outcomes based on multiple categorical and continuous covariates. Specifically, we 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 multi-group 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.

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 $x = 0,1,\ldots, p$. We will take $X = 0$ as reference or control group, and thus $t = 1,\ldots, p$ are the values of the remaining treatment conditions.[1] Furthermore, we will consider a single (unfolded) categorical variable K with j+1 levels and values $k = 0,1,\ldots, j$, and a vector of (latent) variables $\boldsymbol{\xi} = (1, \xi_1, \xi_2, \ldots, \xi_q)$ with $z = 1,\ldots, q$ denoting the (z+1) th element of $\boldsymbol{\xi}$. We use $\boldsymbol{\xi}^{(*)}$ to denote values of $\boldsymbol{\xi}$. For this set of variables, the following parameterization of $E(Y|X,K,\boldsymbol{\xi})$ always holds:

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

$$= \textstyle\sum_{x=0}^{p} \sum_{k=0}^{j} \exp[b_{xk}(\boldsymbol{\xi})] \cdot I_{X=x} \cdot I_{K=k}. \tag{2}$$

In equation (1) the conditional expectation of Y is decomposed into an intercept function g_0 and a conditional effect function g_t for treatment condition t (cf. Mayer, Dietzfelbinger, Rosseel, & Steyer, 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 (2). Note that we use a regression with a logarithmic link function for each combination of $X = x$ and $K = k$, that is, the function $b_{xk}(\xi)$ is group-specific:

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

we use group-specific functions $b_{xk}(\boldsymbol{\xi})$, because the effect computations later in this paper will be based on a multi-group model. The term $E(Y|X = x, K = k, \boldsymbol{\xi})$ denotes a partial conditional expectation. For its definition and further details, see Steyer and Nagel (2017, Section 14.4).

[1] We intentionally use two different indices to refer to the values of the treatment variable X, namely x and t. These will fulfil two distinct purposes later on, that is, we use t to denote the treatment group for which an average or conditional effect (compared to the control group) is computed, while we use x for marginalizing (i.e., summing) over ($X = x$)-conditional distributions. For example, in equation (6) both indices appear and help to distinguish these two aspects.

2.1. Conditional effects function

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

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

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:

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

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 $E(Y|X=x,K,\boldsymbol{\xi})$ reflect the expected outcome given a treatment condition t. For example, $E(Y|X=2,K=1,\xi_1=1)$ denotes the expected count of Y given $K = 1$ for a person who receives training $X = 2$ and has a pre-test score of $\xi_1 = 1$.

In practical settings, a parameterization is needed for $b_{xk}(\boldsymbol{\xi})$, 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, Mayer, and Fiege (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., created by randomized treatment assignment as in the ACTIVE study; see Section 6).

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,

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

where $f_{K,\boldsymbol{\xi}}\left(k,\boldsymbol{\xi}^{(*)}\right)$ denotes the density of the joint distribution of K and $\boldsymbol{\xi}$, 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 randomly sampled person assigned to treatment condition $X = t$ compared to the control group $X = 0$.

2.3. Conditional effects given X and K

In addition to the average effect, finer-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 k of the categorical covariate K and averages over the $(K = k)$-conditional distribution of the continuous covariates,

$$CE_{t0;K=k} = E[g_t(K,\boldsymbol{\xi})|K=k]$$
$$= \int_{\boldsymbol{\xi}^{(*)}} \exp\left[h_{tk}\left(\boldsymbol{\xi}^{(*)}\right)\right] - \exp\left[h_{0k}\left(\xi^{(*)}\right)\right] \cdot f_{\boldsymbol{\xi}|K=k}\left(\boldsymbol{\xi}^{(*)}\right) d\boldsymbol{\xi}^{(*)}.$$

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

$$CE_{t0;K=k} = E[g_t(K,\boldsymbol{\xi})|K=k]$$
$$= \int_{\boldsymbol{\xi}^{(*)}} \exp\left[h_{tk}\left(\boldsymbol{\xi}^{(*)}\right)\right] - \exp\left[h_{0k}\left(\boldsymbol{\xi}^{(*)}\right)\right] \cdot f_{\boldsymbol{\xi}|K=k}\left(\boldsymbol{\xi}^{(*)}\right) d\boldsymbol{\xi}^{(*)}.$$

These effects are of interest when the $(X = x)$-conditional covariate distribution $f_{K,\boldsymbol{\xi}|X=x}$ 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 of X:

$$CE_{t0;X=x,K=k} = E[g_t(K,\boldsymbol{\xi})|X=x,K=k]$$
$$= \int_{\boldsymbol{\xi}^{(*)}} \exp\left[h_{tk}\left(\boldsymbol{\xi}^{(*)}\right)\right] - \exp\left[h_{0k}\left(\boldsymbol{\xi}^{(*)}\right)\right] \cdot f_{\boldsymbol{\xi}|X=x,K=k}\left(\boldsymbol{\xi}^{(*)}\right) d\boldsymbol{\xi}^{(*)}.$$

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.

2.4. Regression parameterization and factorization of the covariates' distribution

Up to 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 $h_{xk}(\boldsymbol{\xi})$ 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 $h_{xk}(\boldsymbol{\xi})$,

$$h_{xk}(\boldsymbol{\xi}) = \alpha'_{xk}\boldsymbol{\xi}, \tag{3}$$

where $\boldsymbol{\alpha}_{xk} = \left(\alpha_{xk0}, \alpha_{xk1}, \ldots, \alpha_{xkq}\right)$ 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 $(X = x, K = k)$.

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

$$f_{K,\xi}\left(k,\xi^{(*)}\right) = \sum_{x=0}^{p} f_{X,K,\xi}\left(x,k,\xi^{(*)}\right) \tag{4}$$

$$= \sum_{x=0}^{p} P(X=x,K=k)\cdot f_{\xi|X=x,K=k}(\xi^{(*)}) \tag{5}$$

In equation (4), the joint distribution of the covariates K and ξ is defined as the marginal distribution of the joint distribution of the treatment variable X, the categorical covariate K, and the continuous covariate ξ. In equation (5), the joint distribution is decomposed into a categorical group part $P(X = x,\ K = k)$ and a group-conditional density $f_{\xi|X=x,K=k}(\xi^{(*)})$ of the continuous variables. This factorization serves two purposes. First, it allows us to identify the conditional densities required for the computation of the aforementioned average and conditional effects, namely $f_{K,\xi}(k,\xi^{(*)})$, $f_{\xi|K=k}(\xi^{(*)})$, $f_{K,\xi|X=x}(k,\xi^{(*)})$, and $f_{\xi|X=x,K=k}(\xi^{(*)})$. Second, we can use a multi-group approach, namely a multi-group structural equation model, to estimate parameters both of the negative binomial regression and of the distribution of the covariates.

The choice of a factorization is without loss of generality for our approach, because an alternative factorization can always be transformed into our factorization. We provide more information on this aspect in the discussion and in Appendix 1.

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

$$\begin{aligned} AE_{t0} &= \sum_{k=0}^{j} \int_{\xi^{(*)}} \left(\exp\left[h_{tk}\left(\xi^{(*)}\right)\right] - \exp\left[h_{0k}\left(\xi^{(*)}\right)\right]\right)\cdot f_{K,\xi}\left(k,\xi^{(*)}\right) d\xi^{(*)} \\ &= \sum_{x=0}^{p}\sum_{k=0}^{j} P(X=x,K=k) \int_{\xi^{(*)}} \left(\exp\left[\alpha'_{tk}\xi^{(*)}\right] - \exp\left[\alpha'_{0k}\xi^{(*)}\right]\right)\cdot f_{\xi|X=x,K=k}\left(\xi^{(*)}\right) d\xi^{(*)}. \end{aligned} \tag{6}$$

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 multi-group structural equation models. As stated before, for a causal interpretation of the treatment effects, the effect function $g_t(K,\xi)$ must be causally unbiased. Causal unbiasedness goes beyond unbiasedness of the estimated parameters and means that there are no unobserved confounders of the treatment effects. It can be achieved, for example, by an (unconditional or conditional) randomized assignment of persons to the treatment groups or by controlling for all confounders. Tools developed in the causal inference literature such as propensity scores (Rosenbaum & Rubin, 1983) can be helpful to meet the 'no unobserved confounders' condition. For an overview of conditions under which causal unbiasedness is achieved, see Steyer et al. (2014). Thus, applied researchers should carefully evaluate whether the parameterization and factorization described here fit their hypotheses and assumptions.

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 ($X = x$, $K = k$)-conditional density $f_{\xi|K,X}$ of $\boldsymbol{\xi}$. In their extension of the moment-based approach, Kiefer and Mayer (2020) suggest assuming ($X = x$)-conditional multivariate normality. Adapting this notion to our distinction between categorical covariates K and continuous covariates $\boldsymbol{\xi}$ and assuming ($X = x, K = k$)-conditional normality, that is, $\boldsymbol{\xi} \sim N_{xk}\left(\mu_{xk}, \sum_{xk}\right)$, the integration part of the average treatment effect in equation (6) can be substituted with moment-generating functions:

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

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. 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.

All of the aforementioned approaches have in common that we use the joint density of the covariates in a maximum likelihood framework to estimate parameters and treatment effects. As these models contain negative binomial regressions, multi-group parts, measurement models and parameters of the non-normal distributions, they can involve many parameters. Thus, large sample sizes may be required for a solution to converge (Jackson, 2003). For smaller samples or cases in which none of the three proposed approaches is feasible, Bayesian modelling and estimation can be an alternative. We discuss Bayesian alternatives in the Discussion section.

4.1. Case 1: Known non-normal joint distribution

The first and probably simplest case is a known alternative non-normal ($X = x$, $K = k$)-conditional joint distribution for $\boldsymbol{\xi}$. 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. Admittedly, the case of knowing a number of suitable non-normal multivariate distributions and their moment-generating function might not be very common, especially for applied researchers.

However, another important scenario for known alternative distributions is ($X = x$, $K = k$)-conditional independence of the covariates $\xi_1, \xi_2, \ldots$, because then the product of univariate distributions corresponds to the marginal distribution. While the ($X = x$, $K = k$)-conditional independence of the covariates might be a strong assumption, it is helpful to see that the joint moment-generating function can be decomposed into the product of the univariate moment-generating functions in this case. For example, if we consider two continuous covariates ξ_1 and ξ_2 with ($X = x$, $K = k$)-conditional independence $\xi_1 \perp \xi_2 | X = x, K = k$, the corresponding density can be decomposed $f_{\xi 1, \xi 2} = f_{\xi 1} \cdot f_{\xi 2}$, and thus the moment-generating function can be written as the product of univariate moment-generating functions,

$$M_{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2 | X = x, K = k}(t_1, t_2) = M_{\boldsymbol{\xi}_1 | X = x, K = k}(t_1) \cdot M_{\boldsymbol{\xi}_2 | X = x, K = k}(t_2),$$

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.

4.2. 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 $\boldsymbol{\xi}$ into a marginal and a conditional distribution, that is,

$$f_{\boldsymbol{\xi} | X = x, K = k} = f_{\boldsymbol{\xi}_m | X = x, K = k} \cdot f_{\boldsymbol{\xi}_c | X = x, K = k, \boldsymbol{\xi}_m = \boldsymbol{\xi}_m^{(*)}},$$

where we specify the marginal distribution of $\boldsymbol{\xi}_m = (\boldsymbol{\xi}_{m_1}, \boldsymbol{\xi}_{m_2}, \ldots)$ with index m_i identifying a subset of covariates, and the conditional distribution of $\boldsymbol{\xi}_c = (\boldsymbol{\xi}_{c1}, \boldsymbol{\xi}_{c2}, \ldots)$ given $\boldsymbol{\xi}_m$ with index c_i identifying the remaining covariates in $\boldsymbol{\xi}$. In general, this factorization makes it possible to simplify the integration part of the average effect in equation (6) as

$$\begin{aligned} &\int_{\boldsymbol{\xi}^{(*)}} \left(\exp\left[\boldsymbol{\alpha}'_{tk} \boldsymbol{\xi}^{(*)} \right] - \exp\left[\boldsymbol{\alpha}'_{0k} \boldsymbol{\xi}^{(*)} \right] \right) \cdot f_{\boldsymbol{\xi} | X = x, K = k}\left(\boldsymbol{\xi}^{(*)} \right) d\boldsymbol{\xi}^{(*)} \\ &= \int_{\boldsymbol{\xi}_m^{(*)}} \left(\exp\left[\boldsymbol{\alpha}'_{tk;m} \boldsymbol{\xi}_m^{(*)} \right] M_{\boldsymbol{\xi}_c | \boldsymbol{\xi}_m = \boldsymbol{\xi}_m^{(*)}}(\boldsymbol{\alpha}_{tk;c}) - \exp\left[\boldsymbol{\alpha}'_{0k;m} \boldsymbol{\xi}_m^{(*)} \right] M_{\boldsymbol{\xi}_c | \boldsymbol{\xi}_m = \boldsymbol{\xi}_m^{(*)}}(\boldsymbol{\alpha}_{0k;c}) \right) \cdot f_{\boldsymbol{\xi}_m | X = x, K = k}\left(\boldsymbol{\xi}_m^{(*)} \right) \cdot d\boldsymbol{\xi}_m^{(*)}. \end{aligned}$$

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 $E(\boldsymbol{\xi}_c | X = x, K = k, \boldsymbol{\xi}_m)$. 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.

4.3. 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, $f_{\boldsymbol{\xi}|X=x,K=k}\left(\boldsymbol{\xi}^{(*)}\right) = \sum_{m=1}^{M} w_m \cdot f_{\boldsymbol{\xi}|X=x,K=k;m}\left(\boldsymbol{\xi}^{(*)}\right)$ where w_m are weights and $\boldsymbol{\xi} \sim N_{xk;m}\left(\boldsymbol{\mu}_{xk;m}, \boldsymbol{\Sigma}_{xk;m}\right)$. In practice, this approximation of the distribution of $\boldsymbol{\xi}$ 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 $\boldsymbol{\mu}_{xk;m}$ and the variance $\sum_{xk;m}$ are allowed to vary among latent classes. This approach has previously been applied in the computation of average and conditional effects in a nonlinear regression setting by Mayer, Umbach, Flunger, and Kelava (2017) and can be estimated using nonlinear structural equation mixture models (Kelava & Brandt, 2014; Kelava, Nagengast, & Brandt, 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 $(X = x, K = k)$-conditional moment-generating function for $\boldsymbol{\xi}$, as we can use a weighted sum of normal moment-generating functions for the integration part from equation (6),

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

which is similar to the original extended moment-based approach by Kiefer and Mayer (2020), with the exception of the summation over the latent classes.

5. Negative binomial multi-group structural equation model

In this section, we introduce the negative binomial multi-group structural equation model (NB-MG-SEM) as a statistical framework for parameter and effect estimation. The NB-MG-SEM provides maximum likelihood 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, and the vector of continuous covariates $\boldsymbol{\xi} = \left(1, \xi_1, \ldots, \xi_q\right)$ containing latent covariates measured by observed variables $\mathbf{z} = (Z_1, Z_2, \ldots, Z_s)$. The NB-MG-SEM consists of (at least) the following parts:

$$\mathbf{z} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{7}$$

$$\mu_Y = \exp\left(\boldsymbol{\alpha}'_{xk}\boldsymbol{\xi}\right), \tag{8}$$

$$\log(n_{xk}) = \kappa_{xk}, \tag{9}$$

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

The NB-MG-SEM presented in equations (7) to (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 between two covariates. We provide detailed information on the maximum likelihood estimation of the NB-MG-SEM for our illustrative example in Appendix 2. For brevity, we present the four main parts and assumptions of the likelihood functions constituting the joint distribution of the variables considered:

$$\begin{gathered}
\eta_{xk} \sim P_{xk}(\exp(\kappa_{xk})), \\
Y \sim NB(\mu_Y, \phi_{xk}), \\
\mathbf{z} \sim N_{xk}(\nu + \Lambda\xi, \Theta_{xk}), \\
\boldsymbol{\xi}\ \textit{depending on case}.
\end{gathered}$$

The joint distribution of the covariates $\boldsymbol{\xi}$ depends on which of the three aforementioned cases is applied. In our illustrative example, we will factorize the distribution of $\boldsymbol{\xi}$ 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. Illustration of the other cases is provided in the online Appendix S1.

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; Tennstedt et al., 2005). 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 pre-test 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. The ACTIVE data subset analysed is publicly available; a link is provided in the online Appendix S1.

In this paper 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 *CountEffects* package. The *CountEffects* package is an implementation of the moment-based approach assuming ($X = x, K = k$)-conditionally multivariate normally distributed covariates, which we extended with functions to estimate all three proposed approaches to deal with non-normal covariates for our illustrative example. In the online Appendix S1, we provide information on how to install *CountEffects* and how to estimate treatment effects based on the multivariate normal assumption and all three cases. Our findings indicate that the three non-normal cases yield similar results for the illustrative example, while the normal case yields differing estimates for some effects. However, in this section we will only describe the estimation and results of the case 2 scenario. 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.

6.1. Sample

The total sample size for our analysis was $N = 2{,}363$. The participants were randomly assigned to one of four conditions: a no-contact control group ($X = 0$, $N = 592$), a memory training ($X = 1, N = 589$), a reasoning training ($X = 2, N = 590$), and a speed of processing training condition ($X = 3, N = 606$). 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 10 days after the last training session. The participants were predominantly female (75.8%, $K = 1$).

6.2. Measures

6.2.1. 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 four-point scale from 0 = never to 3 = 5–7 days. We modelled 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 warnings about the use of parcels, see Marsh, Lüdtke, Nagengast, Morin, and von Davier (2013) and Sterba and Rights (2016).

6.2.2. 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., pre-test 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 7 min to complete the task.

6.3. Model

Our effect analysis was based on a multi-group 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 variable.[2] 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:

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

The latent variable ξ_1 was assumed to have an ($X = x$, $K = k$)-conditional normal distribution

$$\xi_1 \sim N_{xk}(\mu_{xk}, \sigma_{xk})$$

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

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

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 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 3).

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

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

The whole model is displayed in Figure 1. Note that regressions with a logarithmic link function are indicated by curved arrows, in contrast to straight lines for linear regressions.

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 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 ($X = x$, $K = k$)-conditionally

[2] In 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 discuss the issue of estimation difficulties at the end of the paper.

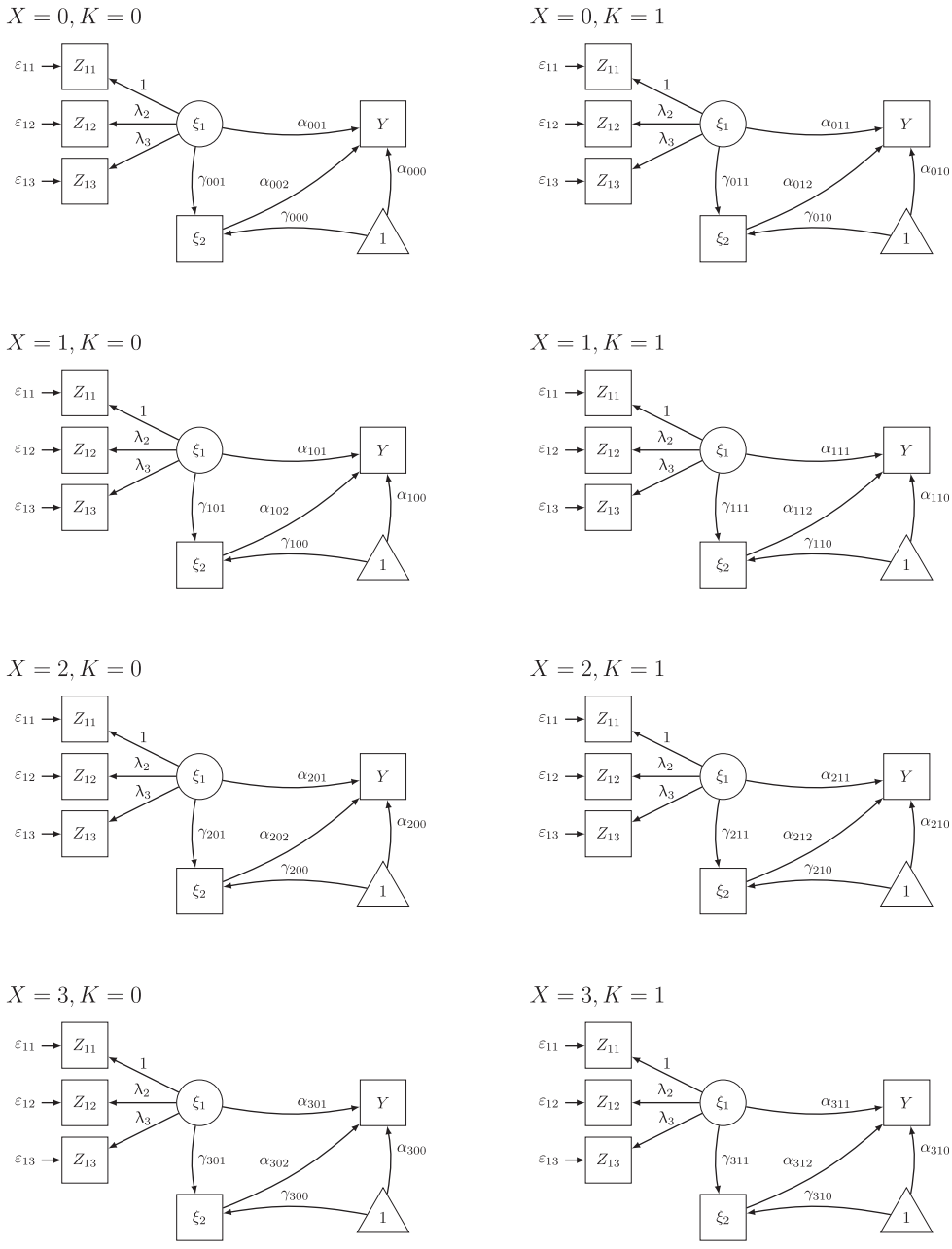

Figure 1. Path diagram for our illustrative example depicting the group invariant measurement models for latent baseline depression ξ_1 and the baseline count of correctly answered items ξ_2, the structural model specifying the regressions with a logarithmic link function (indicated by curved arrows) 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.

marginally normally distributed $\xi_1 \sim N_{xk}(\mu_{xk}, \sigma_{xk})$ and is related to the baseline count of correctly answered items via a Poisson regression with $\xi_2 \sim P_{xk}(\lambda_{xk})$ with $\lambda_{xk} = \exp(\gamma_{xk0} + \gamma_{xk1}\xi_1)$. In this scenario, the average treatment effect can be computed as

$$ATE = \sum_{x=0}^{p}\sum_{k=0}^{j} P(X=x, K=k) \cdot \sum_{b=1}^{H} w_b \left[\exp\left(\alpha_{tk0} + \alpha_{tk1}\xi_{1b}^*\right) \cdot M_{\xi_2|\xi_1, X=x, K=k}(\alpha_{tk2}) \right.$$

$$\left. -\exp\left(\alpha_{0k0} + \alpha_{0k1}\xi_{1b}^*\right) \cdot M_{\xi_2|\xi_1, X=x, K=k}(\alpha_{0k2})\right],$$

where $\xi_{1b}^* = \sqrt{2\sigma_{xk}^2 a_b + \mu_{xk}}$ and a_b, w_b are Gauss–Hermite quadrature points and weights. A detailed derivation of this formula is provided in Appendix 3. 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 as

$$CE_{t0;K=k} = \sum_{x=0}^{p} P(X=x, K=k) \cdot \sum_{b=1}^{H} w_b \left[\exp\left(\alpha_{tk0} + \alpha_{tk1}\xi_{1b}^*\right) \cdot M_{\xi_2|\xi_1, X=x, K=k}(\alpha_{tk2}) \right.$$

$$\left. -\exp\left(\alpha_{0k0} + \alpha_{0k1}\xi_{1b}^*\right) M_{\xi_2|\xi_1, X=x, K=k}(\alpha_{0k2})\right],$$

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

Standard errors for the estimated treatment effects can be derived using the delta method (cf. Boos & Stefanski, 2013, p. 237).[3] In our analysis, we computed symmetric confidence intervals based on the standard errors, because the estimated treatment effects are asymptotically normally distributed and our sample was comparatively large. However, in smaller samples confidence intervals based on, for example, a bootstrap approach are recommended.

6.5. Results

It is currently 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 nonlinearity.

6.5.1. Model parameters

The maximum likelihood estimates for the model parameters are given in Table 1. We will not discuss them in detail here, but rather present some notable findings to illustrate the interpretation of these parameters.

[3] For a historical side note on the question of who invented the delta method, see Ver Hoef (2012).

Table 1. Maximum likelihood results for group-specific model parameters

Parameter	Estimate	*SE*	*p*	Parameter	Estimate	*SE*	*p*
$X = 0, K = 0$				$X = 0, K = 0$			
Estimates for $E(Y\|X=0,K=0,\xi_1,\xi_2)$				Estimates for $E(Y\|X=1,K=1,\xi_1,\xi_2)$			
α_{000}	1.324	0.103	<.001	α_{010}	1.218	0.061	<.001
α_{001}	−0.048	0.041	.241	α_{011}	-0.034	0.018	.061
α_{002}	0.099	0.011	<.001	α_{012}	0.103	0.007	<.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}	1.970	0.059	<.001	γ_{010}	1.904	0.035	<.001
γ_{001}	0.148	0.049	.002	γ_{011}	-0.109	0.020	<.001
Estimates for $\xi_1\|X=0,K=0$				Estimates for $\xi_1\|X=0,K=0$			
μ_{00}	1.113	0.087	<.001	μ_{01}	1.571	0.072	<.001
σ^2_{00}	0.792	0.128	<.001	σ^2_{01}	1.629	0.149	<.001
$X = 1, K = 0$				$X = 1, K = 1$			
Estimates for $E(Y\|X=0,K=0,\xi_1,\xi_2)$				Estimates for $E(Y\|X=1,K=1,\xi_1,\xi_2)$			
α_{100}	1.141	0.110	<.001	α_{110}	1.256	0.057	<.001
α_{101}	0.016	0.031	.602	α_{111}	-0.037	0.016	.019
α_{102}	0.111	0.013	<.001	α_{112}	0.102	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}	1.938	0.053	<.001	γ_{110}	1.890	0.030	<.001
γ_{101}	-0.057	0.034	.094	γ_{111}	-0.077	0.017	<.001
Estimates for $\xi_1\|X=1,K=0$				Estimates for $\xi_1\|X=1,K=1$			
μ_{10}	1.296	0.109	<.001	μ_{11}	1.440	0.074	<.001
σ^2_{10}	1.298	0.201	<.001	σ^2_{11}	1.955	0.176	<.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}	1.335	0.103	<.001	α_{210}	1.340	0.053	<.001
α_{201}	0.004	0.026	.890	α_{211}	-0.013	0.014	.345
α_{202}	0.102	0.011	<.001	α_{212}	0.099	0.006	<.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}	2.002	0.052	<.001	γ_{210}	1.785	0.033	<.001
γ_{201}	-0.098	0.031	.001	γ_{211}	-0.024	0.016	.145
Estimates for $\xi_1\|X=2,K=0$				Estimates for $\xi_1\|X=2,K=1$			
μ_{20}	1.411	0.127	<.001	μ_{21}	1.641	0.074	<.001
α_{20}	1.879	0.303	<.001	α_{21}	1.879	0.171	<.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}	1.274	0.101	<.001	α_{310}	1.210	0.057	<.001
α_{301}	0.016	0.026	.532	α_{311}	-0.009	0.017	.581
α_{302}	0.095	0.012	<.001	α_{312}	0.107	0.007	<.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}	1.936	0.051	<.001	γ_{310}	1.823	0.033	<.001
γ_{301}	-0.081	0.030	.007	γ_{311}	-0.052	0.018	.005
Estimates for $\xi_1\|X=3,K=0$				Estimates for $\xi_1\|X=3,K=1$			
μ_{30}	1.416	0.126	<.001	μ_{31}	1.526	0.067	<.001
α_{30}	1.861	0.277	<.001	α_{31}	1.515	0.137	<.001

Pre-test depression had no significant effect on the post-test count of correctly answered items in either group (e.g., $\alpha_{001} = 0.048$, 95% confidence interval (CI) [−0.129, 0.033]), except for female participants receiving the memory training ($\alpha_{111} = -0.037$,

95% CI [−0.069, 0.006]). Furthermore, pre-test depression was significantly negatively related to the pre-test count of correctly answered items in most groups (e.g., $\gamma_{011} = -0.109$, 95% CI [−0.148, 0.070], $\gamma_{201} = -0.098$, 95% CI [−0.758, 0.038], and $\gamma_{301} = -0.081$, 95% CI [−0.139, 0.022]). For example, for male participants in the reasoning training, each one-unit change in pre-test depression was linked to a 9% decrease in correctly answered items at baseline (i.e., γ_{201}, $\exp(-0.098) = -0.91$). In two groups, this relationship was not significant ($\gamma_{101} = -0.057$, 95% CI [−0.123, 0.010] and $\gamma_{211} = -0.024$, 95% CI [−0.056, 0.008]).

Pre-test 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 an 11% increase in correctly answered items at post-test (i.e., α_{112}, $\exp(0.102) = 1.11$). The intercept coefficients reflect the expected post-test 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., $\alpha_{200} = 1.335$), the expected count is $\hat{E}(Y|X = 2, K = 0, \xi_1 = 0, \xi_2 = 0) = \exp(1.335) = 3.80$.

6.5.2. Average effects

An overview of all average and conditional treatment effects estimated for our illustrative example is given in Table 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 training might not necessarily affect reasoning performance. In line with this, the average treatment effect of the memory training ($\widehat{AE}_{10} = 0.084$, 95% CI [−0.203, 0.372], ES = 0.030) and the average treatment effect of the speed of processing training ($\widehat{AE}_{30} = 0.237$, 95% CI [−0.052, 0.525], ES = 0.083) on the count of correctly answered items were not significant. The average treatment effect of the reasoning training, however, was significant ($\widehat{AE}_{20} = 0.783$, 95% CI [0.487, 1.080], ES = 0.275), that is, participants in this condition correctly answered 0.849 items more on average compared to baseline.

6.5.3. Conditional effects given (K = k)

The conditional treatment effects given a gender are given in detail in Table 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 ($\widehat{CE}_{10;K=0} = -0.152$, 95% CI [−0.750, 0.446], ES = −0.053) and a positive treatment effect for female participants $K = 1$ ($\widehat{CE}_{10;K=1} = 0.161$, 95% CI [−0.167, 0.489], ES = 0.057). 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 ($\widehat{CE}_{20;K=0} = 0.819$, 95% CI [0.482, 1.156], ES = 0.288) was higher than the conditional effect for male participants ($\widehat{CE}_{20;K=0} = 0.674$, 95% CI [0.052, 1.297], ES = 0.237).

While the average treatment effect of the speed of processing training was not significant, we found a slightly non-significant conditional effect for female participants ($\widehat{CE}_{30;K=1} = 0.325$, 95% CI [−0.002, 0.652], ES = 0.114). Conversely, for male participants, the speed of processing training did not yield a significant effect

Table 2. Estimated average and conditional effects of memory training ($X = 1$ versus $X = 0$), reasoning training ($X = 2$ versus $X = 0$), and speed of processing training ($X = 3$ versus $X = 0$) 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 $\widehat{CE}_{10;X=x,K=k}$			
$\widehat{AE}_{10}$	0.084	0.147	0.030	$\widehat{CE}_{10;X=0,K=0}$	−0.238	0.302	−0.084
Conditional effects $\widehat{CE}_{10;X=x}$				$\widehat{CE}_{10;X=0,K=1}$	0.161	0.166	0.056
$\widehat{CE}_{10;X=0}$	0.053	0.147	0.019	$\widehat{CE}_{10;X=1,K=0}$	−0.130	0.314	−0.046
$\widehat{CE}_{10;X=1}$	0.094	0.151	0.033	$\widehat{CE}_{10;X=1,K=1}$	0.167	0.171	0.059
$\widehat{CE}_{10;X=2}$	0.093	0.147	0.033	$\widehat{CE}_{10;X=2,K=0}$	−0.109	0.321	−0.038
$\widehat{CE}_{10;X=3}$	0.095	0.147	0.034	$\widehat{CE}_{10;X=2,K=1}$	0.157	0.167	0.055
Conditional effects $\widehat{CE}_{10;K=k}$				$\widehat{CE}_{10;X=3,K=0}$	−0.119	0.309	−0.042
$\widehat{CE}_{10;K=0}$	−0.152	0.305	−0.053	$\widehat{CE}_{10;X=3,K=1}$	0.160	0.167	0.056
$\widehat{CE}_{10;K=1}$	0.161	0.167	0.057				
Estimated average and conditional effects of reasoning training							
Average effect $\widehat{AE}_{20}$				Conditional effects $\widehat{CE}_{20;X=x,K=k}\widehat{CE}_{20;X=x,K=k}$			
$\widehat{AE}_{20}$	0.784	0.151	0.275	$\widehat{CE}_{20;X=0,K=0}$	0.598	0.317	0.210
Conditional effects $\widehat{CE}_{20;X=x}$				$\widehat{CE}_{20;X=0,K=1}$	0.814	0.171	0.286
$\widehat{CE}_{20;X=0}$	0.756	0.152	0.266	$\widehat{CE}_{20;X=1,K=0}$	0.697	0.325	0.245
$\widehat{CE}_{20;X=1}$	0.786	0.156	0.276	$\widehat{CE}_{20;X=1,K=1}$	0.814	0.178	0.286
$\widehat{CE}_{20;X=2}$	0.797	0.152	0.280	$\widehat{CE}_{20;X=2,K=0}$	0.715	0.329	0.251
$\widehat{CE}_{20;X=3}$	0.789	0.151	0.277	$\widehat{CE}_{20;X=2,K=1}$	0.831	0.171	0.292
Conditional effects $\widehat{CE}_{20;K=k}$				$\widehat{CE}_{20;X=3,K=0}$	0.697	0.319	0.245
$\widehat{CE}_{20;K=0}$	0.674	0.318	0.237	$\widehat{CE}_{20;X=3,K=1}$	0.817	0.171	0.287
$\widehat{CE}_{20;K=1}$	0.819	0.172	0.288				
Estimated average and conditional effects of speed of processing training							
Average effect $\widehat{AE}_{30}$				Conditional effects $\widehat{CE}_{30;X=x,K=k}$			
$\widehat{AE}_{30}$	0.237	0.147	0.083	$\widehat{CE}_{30;X=0,K=0}$	−0.102	0.308	−0.036
Conditional effects $\widehat{CE}_{30;X=x}$				$\widehat{CE}_{30;X=0,K=1}$	0.321	0.166	0.113
$\widehat{CE}_{30;X=0}$	0.207	0.147	0.073	$\widehat{CE}_{30;X=1,K=0}$	−0.025	0.320	−0.009
$\widehat{CE}_{30;X=1}$	0.234	0.152	0.082	$\widehat{CE}_{30;X=1,K=1}$	0.317	0.172	0.111
$\widehat{CE}_{30;X=2}$	0.254	0.148	0.089	$\widehat{CE}_{30;X=2,K=0}$	−0.015	0.326	−0.005
$\widehat{CE}_{30;X=3}$	0.250	0.147	0.088	$\widehat{CE}_{30;X=2,K=1}$	0.340	0.167	0.120
Conditional effects $\widehat{CE}_{30;K=k}$				$\widehat{CE}_{30;X=3,K=0}$	0.011	0.314	0.004
$\widehat{CE}_{30;K=0}$	−0.035	0.310	−0.012	$\widehat{CE}_{30;X=3,K=1}$	0.322	0.166	0.113
$\widehat{CE}_{30;K=1}$	0.325	0.167	0.114				

($\widehat{CE}_{30;K=0} = -0.035$, 95% CI [−0.750, 0.446], ES = −0.053). This differential effect illustrates why examining conditional effects can be crucial when evaluating a treatment.

6.5.4. Conditional effects given (X = x) and (X = x, K = k)

The conditional treatment effects given a treatment condition (and gender) are given in detail in Table 2. As the ACTIVE study was a randomized controlled trial, the ($X = x$, $K = k$)-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 ($\widehat{CE}_{10;X=0} = 0.053$, 95% CI [−0.234, 0.340], ES = 0.019) 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 had instead 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.

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, 2020) in three respects. First, we presented four ways to account for non-normal covariates: (1) for categorical covariates, applying a saturated model; for continuous covariates, we suggested either (2) the use of an alternative, known non-normal joint distribution (e.g., skew-normal distribution), (3) a plausible factorization into marginal and conditional distributions, or (4) 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 finer-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 as covariates the baseline count of correctly answered items (non-normal covariate), baseline depression (latent covariate), and gender (categorical covariate). The corresponding negative binomial multi-group structural equation model and the average and conditional treatment effect estimations were carried out in R. We have 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 covariate values as random and not predetermined by the experimenter. Ignoring this randomness would lead to underestimation of standard errors and, thus, inflated Type I error rates and decreased power, which has previously been shown for stochastic covariates (Li, McLouth, & Delaney, 2020; Liu, West, Levy, & Aiken, 2017) and stochastic group sizes (Mayer & Thoemmes, 2019). In addition, the moment-based approach allows accounting for measurement error in covariates. For an overview of consequences of measurement error in nonlinear regression models, see Carroll, Ruppert, Stefanski, and Crainiceanu (2010). However, in the social and health sciences, observed variables in real data sets often deviate from the normal distribution (Bono et al., 2017; Micceri, 1989). While Kiefer and Mayer (2020) 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.

7.1. Limitations and further research

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

In equation (3), we assumed a linear parameterization for the predictor function $h_{xk}(\xi)$. 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 nonlinear predictor functions might be of substantial interest. For example, Mayer et al. (2017) investigated the effectiveness of autonomy support by ninth-grade teachers in reducing students' state of boredom. They hypothesized a quadratic relationship between boredom and self-efficacy, which means that the treatment could be most effective for medium values of self-efficacy and less for both extremes. Similarly, Liu and West (2015) use trigonometric terms to predict cyclic patterns in daily report (count) data. The inclusion of quadratic or other nonlinear terms might also be of interest 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 nonlinear 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 $(X = x,\ K = k)$-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 $(Z = z)$-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 rewritten in our factorization, that is, a marginal distribution of the binary treatment and an $(X = x)$-conditional distribution of Z (for a derivation, see Appendix 1).

We used maximum likelihood estimation to obtain parameter estimates for the negative binomial multi-group 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 address non-convergence issues. First, a pragmatic workaround 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 modelling framework can be applied. The Bayesian approach can typically handle complex models much easier than maximum likelihood estimation (Merkle & Rosseel, 2015). For an overview of Bayesian estimation of structural equation models based on finite-mixture distributions or distributions based on the exponential family, see Lee (2007, Chapters 11 and 13). 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 maximum likelihood estimation. However, if the factor scores depend on a distributional

assumption, it would be wise to use the same distributional assumption within the moment-based approach.

In our illustrative example, we examined the effectiveness of cognitive training using a pre–post design. The ACTIVE study also contains several follow-up measurements allowing long-term effectiveness of the cognitive training to be investigated. 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, Rose, Steyer, & West, 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 non-randomized 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 Reichhardt (2019).

References

Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715–726. https://doi.org/10.1093/biomet/83.4.715

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., . . . for the ACTIVE Study Group (2002). Effects of cognitive training interventions with older adults. *Journal of the American Medical Association*, *288*, 2271. https://doi.org/10.1001/jama.288.18.2271

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.

Bono, R., Blanca, M. J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. *Frontiers in Psychology*, *8*, 1602. https://doi.org/10.3389/fpsyg.2017.01602

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: Theory and methods*. New York, NY: Springer.

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.

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*, *39*(1), 1–38.

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*, 580. https://doi.org/10.3389/fpsyg.2018.00580

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*, *13*, 31–38. https://doi.org/10.1027/1614-2241/a000130

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. https://doi.org/10.1016/j.jsat.2017.10.012

Geneletti, S., & Dawid, A. P. (2011). Defining and identifying the effect of treatment on the treated. *Causality in the sciences* (pp. 728–749). Oxford, UK: Oxford University Press.

Greene, W. (2007). *Econometric analysis*. Upper Saddle River, NJ: Prentice Hall.

Hittner, J. B., Owens, E. C., & Swickert, R. J. (2016). Influence of social settings on risky sexual behavior. *SAGE Open*, *6*(1), 215824401662918. https://doi.org/10.1177/2158244016629187

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 205–223. https://doi.org/10.1207/s15328007sem0802_3

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.

Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., . . . Kleinman, K. (2001). ACTIVE. *Controlled Clinical Trials*, *22*, 453–479. https://doi.org/10.1016/S0197-2456(01)00139-8

Kelava, A., & Brandt, H. (2014). A general nonlinear multilevel structural equation mixture model. *Frontiers in Psychology*, *5*(748), 748. https://doi.org/10.3389/fpsyg.2014.00748

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. https://doi.org/10.1080/10705511.2014.915379

Kiefer, C., & Mayer, A. (2019). Average effects based on regressions with a logarithmic link function: A new approach with stochastic covariates. *Psychometrika*, *84*, 422–446. https://doi.org/10.1007/s11336-018-09654-1

Kiefer, C., & Mayer, A. (2020). Accounting for latent covariates in average effects from count regressions. *Multivariate Behavorial Research*, https://doi.org/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*, 189–221.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, UK: Wiley.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling*, *12*, 1–27.

Li, L., McLouth, C. J., & Delaney, H. D. (2020). 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*, *55*, 926–940. https://doi.org/10.1080/00273171.2019.1693953

Liu, Y., & West, S. G. (2015). Weekly cycles in daily report data: An overlooked issue. *Journal of Personality*, *84*, 560–579. https://doi.org/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. https://doi.org/10.1080/00273171.2017.1309261

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J. S., & von Davier, M. (2013). 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*, 257–284. https://doi.org/10.1037/a0032773

Mayer, A. (2019). Causal effects based on latent variable models. *Methodology*, *15*, 15–28. https://doi.org/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. https://doi.org/10.1080/00273171.2016.1151334

Mayer, A., Steyer, R., & Mueller, H. (2012). A general approach to defining latent growth components. *Structural Equation Modeling*, *19*, 513–533. https://doi.org/10.1080/10705511.2012.713242

Mayer, A., & Thoemmes, F. (2019). Analysis of variance models with stochastic group weights. *Multivariate Behavioral Research*, *54*, 542–554. https://doi.org/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*, 425–442. https://doi.org/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*, 556–570. https://doi.org/10.1080/10705511.2016.1273780

Mazerolle, L., Bennett, S., Antrobus, E., Cardwell, S. M., Eggins, E., & Piquero, A. R. (2019). 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*, *35*, 663–689. https://doi.org/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.

McCullagh, P., & Nelder, J. A. (1996). *Generalized linear models* (2nd ed.). London, UK: Chapman & Hall.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons.

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1-30. https://doi.org/10.18637/jss.v085.i04

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th edn). Los Angeles, CA: Author.

Nusser, L., & Weinert, S. (2017). Appropriate test-taking instructions for students with special educational needs. *Journal of Cognitive Education and Psychology*, *16*, 227–240. https://doi.org/10.1891/1945-8959.16.3.227

R Core Team (2018). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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.

Reichhardt, C. (2019). *Quasi-experimentation: A guide to design and analysis*. New York, NY: Guilford.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

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*, 317–333. https://doi.org/10.1037/met0000013

Schaumberg, R. L., & Flynn, F. J. (2017). Clarifying the link between job satisfaction and absenteeism: The role of guilt proneness. *Journal of Applied Psychology*, *102*, 982. https://doi.org/10.1037/apl0000208

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. https://doi.org/10.2196/14602

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*, 296–313. https://doi.org/10.1080/00273171.2016.1144502

Steyer, R., Mayer, A., & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 606–631). Dordrecht, The Netherlands: Springer. https://doi.org/10.1007/978-94-007-0753-5_295

Steyer, R., & Nagel, W. (2017). *Probability and conditional expectation*. Chichester, UK: John Wiley & Sons.

Tennstedt, S., Morris, J., Unverzagt, F., Rebok, G., Willis, S., Ball, K., & Marsiske, M. (2005). ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly), United States, 1999–2001. *Interuniversity Consortium for Political and Social Research*, https://doi.org/10.3886/ICPSR04248.V3

Ver Hoef, J. M. (2012). Who invented the delta method? *American Statistician*, *66*, 124–127. https://doi.org/10.1080/00031305.2012.687494

Received 21 April 2020; revised version received 12 January 2021

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1 Instructions on estimation and comparison of our illustrative example using an R package.

Appendix 1:

Factorization of the joint distribution

In equations (4) and (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 $f_{A,B}(a,b)$. Then

$$f_{A,B}(a,b) = f_A(a) \cdot f_{B|A=\alpha}(b) = f_B(b) \cdot f_{A|B=b}(a),$$
$$f_A(a) = \int_b f_{A,B}(a,b)db.$$

If the marginal distribution $f_B(b)$ and the conditional distribution $f_{A|B=b}(a)$ are known, the corresponding distributions $f_A(a)$ and $f_{B|A=a}(b)$ can directly be computed with

$$f_A(a) = \int_b f_B(b) \cdot f_{A|B=b}(a)db,$$
$$f_{B|A=a}(b) = \frac{f_B(b) \cdot f_{A|B=b}(a)}{f_A(a)} = \frac{f_B(b) \cdot f_{A|B=b}(a)}{\int_b f_B(b) \cdot f_{A|B=b}(a)db}.$$

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) and (5), we now consider

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

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

$$P(X=x|Z=z)=\left[\frac{\exp(\upsilon_0+\upsilon_1 z)}{1+\exp(\upsilon_0+\upsilon_1 z)}\right]^{x}\cdot\left[1-\frac{\exp(\upsilon_0+\upsilon_1 z)}{1+\exp(\upsilon_0+\upsilon_1 z)}\right]^{(1-x)},$$

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 (5) can be obtained from these known marginal and conditional distributions by computing

$$P(X=x)=\int_z f_Z(z)\cdot P(X=x|Z=z)dz,$$
$$f_{Z|X=x}=\frac{f_Z(z)\cdot P(X=x|Z=z)}{\int_z f_Z(z)\cdot P(X=x|Z=z)dz}.$$

Note that these transformations are generally required for effect computation only. The maximum likelihood estimation would, nevertheless, be based on $f_Z(Z)$ and $P(X = x|Z = z)$ in this example. In contrast, when trying to estimate the likelihood function based on $P(X = x)$ and $f_{Z|X=x}$, it might be tempting to estimate $P(X = x)$ directly instead of $\int_z f_Z(Z)\cdot P(X=x|Z=z)dz$. 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,υ_0, and υ_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 at its starting value, and thus would not yield a trustworthy estimate. The likelihood function should only contain parameters μ_Z, σ_Z^2,υ_0, and υ_1. Corresponding parameters from our factorization and their standard errors can then be computed for effect estimation using the delta method.

Appendix 2:

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 four 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_1. We want to estimate a statistical model as presented in equations (7) to (9) using maximum likelihood techniques. Assuming a sample of N independently and identically distributed observations, the complete-data likelihood is given by

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

where y_i, x_i, $\boldsymbol{k}_i$, $\boldsymbol{z}_i$ $\xi_1^{(*)}$ are the ith observed values of the aforementioned variables. As we do not have observed values $\xi_1^{(*)}$ i.e., latent covariate), we actually have to optimize the observed-data likelihood:

$$\mathscr{L}(\theta|Y,X,K,\mathbf{z},\xi_2) = \prod_{i=1}^{N} f\left(y_i,\boldsymbol{x}_i,\boldsymbol{k}_i,z_i,\xi_{2i}^{(*)}|\boldsymbol{\theta}\right)$$
$$= \prod_{i=1}^{N} \int_{\xi_1} f\left(y_i,\boldsymbol{x}_i,\boldsymbol{k}_i,z_i,\xi_i^{(*)}|\boldsymbol{\theta}\right) d\xi_1^{(*)}.$$

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 is beyond the scope of this paper.

The joint density $f(y_i,\boldsymbol{x}_i,\boldsymbol{k}_i,\mathbf{z}_i,\ \boldsymbol{\xi}_i^{(*)}|\boldsymbol{\theta})$can be factorized into five conditional densities:

$$f(y_i,\boldsymbol{x}_i,\boldsymbol{k}_i,\mathbf{z}_i,\ \boldsymbol{\xi}_i^{(*)}|\boldsymbol{\theta}) = f(n_{x_ik_i}|\boldsymbol{\theta}_1)$$

$$f(y_i|\boldsymbol{\theta}_2,\boldsymbol{x}_i,\boldsymbol{k}_i,\ \boldsymbol{\xi}_i^{(*)})$$

$$f(\mathbf{z}_i|\boldsymbol{\theta}_3,\boldsymbol{x}_i,\boldsymbol{k}_i,\ \xi_{1i}^{(*)})$$

$$f(\ \xi_{2i}^{(*)}|\boldsymbol{\theta}_4,\boldsymbol{x}_i,\boldsymbol{k}_i,\xi_{1i})$$

$$f(\ \xi_{1i}^{(*)}|\boldsymbol{\theta}_5,\boldsymbol{x}_i,\boldsymbol{k}_i).$$

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:

$$\boldsymbol{\theta}_1 = (\kappa_{00},\cdots,\kappa_{pj}),$$

$$\boldsymbol{\theta}_2 = (\boldsymbol{\alpha}_{00},\cdots,\boldsymbol{\alpha}_{pj}),$$

$$\boldsymbol{\theta}_3 = (\boldsymbol{\nu},\boldsymbol{\Lambda},\boldsymbol{\Theta}_{00},\cdots,\boldsymbol{\Theta}_{pj}),$$

$$\boldsymbol{\theta}_4 = (\boldsymbol{\gamma}_{00},\cdots,\boldsymbol{\gamma}_{pj}),$$

$$\boldsymbol{\theta}_5 = (\mu_{00},\cdots,\mu_{pj},\sigma_{00}^2,\cdots,\sigma_{pj}^2).$$

In the NB-MG-SEM, the conditional densities are defined by the model assumptions. Here, $f(n_{x_ik_i}|\boldsymbol{\theta}_1)$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 $f(y_i|\boldsymbol{\theta}_2,\boldsymbol{x}_i,\boldsymbol{k}_i,\ \boldsymbol{\xi}_i^{(*)})$, the group-specific and $\xi_1 = \ \xi_{1i}^{(*)}$-conditional distribution of the observed indicators $\mathbf{z}$ is

given by $f(\mathbf{z}_i|\boldsymbol{\theta}_3, x_i, k_i, \xi_{1i}^{(*)})$, 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

$$f(n_{x_i k_i}|\boldsymbol{\theta}_1) = \frac{\exp(\kappa_{xk}^{n_{x_i k_i}})}{n_{x_i k_i}!}\exp(-\exp[\kappa_{xk}]),$$

$$f\left(y_i|\boldsymbol{\theta}_2, x_i, k_i, \xi_i^{(*)}\right) = \frac{\exp\left(\boldsymbol{\alpha}'_{xk}\boldsymbol{\xi}_i^{(*)}\right)^{y_i}}{y_i!}\exp\left(-\exp[\boldsymbol{\alpha}'_{xk}\boldsymbol{\xi}_i^{(*)}]\right),$$

$$f\left(z_i|\boldsymbol{\theta}_3, x_i, k_i, \xi_{1i}^{(*)}\right) = \frac{1}{\sqrt{(2\pi)^3|\Theta_{xk}|}}\exp\left(-\frac{1}{2}(z_i - \xi_{1i}^{(*)})'\Theta_{xk}^{-1}(z_i - \xi_{1i}^{(*)})\right),$$

$$f\left(\xi_{2i}^{(*)}|\boldsymbol{\theta}_4, x_i, k_i, \xi_{1i}^{(*)}\right) = \frac{\exp\left(\boldsymbol{\gamma}'_{xk}(1, \xi_{1i}^{(*)})'\right)^{\xi_{2i}^{(*)}}}{\xi_{2i}^{(*)}!}\exp\left(-\exp\left[\boldsymbol{\gamma}'_{xk}(1, \xi_{1i}^{(*)})'\right]\right),$$

$$f\left(\xi_{1i}^{(*)}|\boldsymbol{\theta}_5, x_i, k_i\right) = \frac{1}{\sqrt{2\pi\sigma_{xk}^2}}\exp\left(-\frac{\left(\xi_{1i}^{(*)} - \mu_{xk}\right)^2}{2\sigma_{xk}^2}\right).$$

As the group weights do not depend on the latent covariate ξ_2, we can simplify the observed-data density as follows:

$$\begin{aligned} f\left(y_i, x_i, k_i, z_i, \xi_{2i}^{(*)}|\theta\right) &= \int_{\xi_1} f\left(y_i, x_i, k_i, z_i, \xi_i^{(*)}|\theta\right) d\xi_1^{(*)} \\ &= f(\eta_{x_i k_i}|\theta) \cdot \int_{\xi_1} f\left(y_i, z_i, \xi^{(*)}|\theta, x_i, k_i\right) d\xi_1^{(*)}. \end{aligned}$$

Hence, the corresponding log-likelihood is

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

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

$$\int_{\xi_1^{(*)}} f\left(y_i, z_i, \xi_i^{(*)} | \boldsymbol{\theta}, x_i, k_i\right) d\xi_1^{(*)}$$
$$\approx \sum_{b=1}^{H} w_b \cdot f\left(y_i | \boldsymbol{\theta}_2, x_i, k_i, {\xi^{(*)}}^*_{1b}, \xi_{2i}^{(*)}\right) \cdot f\left(z_i | \boldsymbol{\theta}_3, x_i, k_i, {\xi^{(*)}}^*_{1b}\right) \cdot f\left(\xi_{2i}^{(*)} | \boldsymbol{\theta}_4, x_i, k_i, {\xi^{(*)}}^*_{1b}\right),$$

where $\xi^*_{1b} = \sqrt{2\sigma^2_{xk}} a_b + \mu_{xk}$ and a_b, w_b are Gauss–Hermite quadrature points and weights.

Appendix 3:

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 four 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 $\xi_1 | X = x, K = k \tilde{\sim} \mathcal{N}_{xk}(\mu_{xk}, \sigma_{xk})$ and is related to the baseline count of correctly answered items via a Poisson regression with $\xi_2 | X = x, K = k, \xi_1 \tilde{\sim} P_{xk}(\lambda_{xk})$ where $\lambda_{xk} = \exp(\gamma_{xk0} + \gamma_{xk1}\xi_1)$.

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

$$ATE = \sum_{x=0}^{p} \sum_{k=0}^{j} P(X = x, K = k) \cdot \int_{\xi_1^{(*)}} \int_{\xi_2^{(*)}} \left(\exp\left[\boldsymbol{\alpha}'_{0k} \boldsymbol{\xi}\right]\right)$$
$$\cdot f_{\xi_1 | K,X}\left(\xi_1^{(*)} | k, x\right) f_{\xi_2 | K,X}\left(\xi_2^{(*)} | \xi_1^{(*)} k, x\right) d\xi_1^{(*)} d\xi_2^{(*)}$$

the integration part can be rewritten as a difference of integrals,

$$= \sum_{x=0}^{p} \sum_{k=0}^{j} P(X = x, K = k)$$
$$\cdot \left(\int_{\xi_1^{(*)}} \int_{\xi_2^{(*)}} \exp\left[\boldsymbol{\alpha}'_{tk} \boldsymbol{\xi}\right] \cdot f_{\xi_1,K,X}\left(\xi_1^{(*)} | k, x\right) f_{\xi_2 | \xi_1,K,X}\left(\xi_2^{(*)} | \xi_1^{(*)}, k, x\right) d\xi_1^{(*)} d\xi_2^{(*)} \right)$$
$$\left(- \int_{\xi_1^{(*)}} \int_{\xi_2^{(*)}} \exp\left[\boldsymbol{\alpha}'_{0k} \boldsymbol{\xi}\right] \cdot f_{\xi_1,K,X}\left(\xi_1^{(*)} | k, x\right) f_{\xi_2 | \xi_1,K,X}\left(\xi_2^{(*)} | \xi_1^{(*)}, k, x\right) d\xi_1^{(*)} d\xi_2^{(*)} \right),$$

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,

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

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

$$\int_{\xi_2^{(*)}} \exp[\alpha_{xk2}\xi_2] \cdot f_{\xi_2|\xi_1,K,X}\left(\xi_2^{(*)}|\xi_1^{(*)},\boldsymbol{k},\boldsymbol{x}\right) d\xi_2^{(*)} = E(\exp[\alpha_{xk2}\xi_2]|X,K,\xi_1)$$
$$= M_{\xi_2|\xi_1,K,X}(\alpha_{xk2}).$$

Consequently, the computation of the average treatment effect simplifies to

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

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

$$\sum_{x=0}^{p}\sum_{k=0}^{j} P(X=\boldsymbol{x},K=\boldsymbol{k})$$
$$\cdot\int_{\xi_1^{(*)}} \left[\exp[\alpha_{tk0}+\alpha_{tk1}\xi_1] M_{\xi_2|\xi_1,K,X}(\alpha_{tk2}) - \exp[\alpha_{0k0}+\alpha_{0k1}\xi_1] M_{\xi_2|\xi_1,K,X}(\alpha_{0k2})\right] f_{\xi_1|K,X}\left(\xi_1^{(*)}|\boldsymbol{k},\boldsymbol{x}\right) d\xi_1^{(*),}$$

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

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

where $\xi_{1b}^* = \sqrt{2\sigma_{xk}^2}a_b + \mu_{xk}$ and a_b, w_b are Gauss–Hermite quadrature points and weights.