Lüdtke, Oliver; Robitzsch, Alexander

Commentary regarding the section "Modelling the effectiveness of teaching quality". Methodological challenges in assessing the causal effects of teaching

Praetorius, Anna-Katharina [Hrsg.]; Grünkorn, Juliane [Hrsg.]; Klieme, Eckhard [Hrsg.]: Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen. 1. Auflage. Weinheim; Basel : Beltz Juventa 2020, S. 210-222. - (Zeitschrift für Pädagogik, Beiheft; 66)

66. Beiheft April 2020

ZEITSCHRIFT FÜR PÄDAGOGIK

Empirische Forschung zu Unterrichtsqualität. Theoretische Grundfragen und quantitative Modellierungen

BELTZ JUVENTA

Zeitschrift für Pädagogik · 66. Beiheft

Zeitschrift für Pädagogik · 66. Beiheft

Empirische Forschung zu Unterrichtsqualität

Theoretische Grundfragen und quantitative Modellierungen

Herausgegeben von
Anna-Katharina Praetorius, Juliane Grünkorn
und Eckhard Klieme

BELTZ JUVENTA

ISSN: 0514-2717
ISBN 978-3-7799-3534-6 Print
ISBN 978-3-7799-3535-3 E-Book (PDF)
Bestellnummer: 443534

1. Auflage 2020

Herstellung: Hannelore Molitor
Satz: text plus form, Dresden
Druck und Bindung: Beltz Grafische Betriebe, Bad Langensalza
Printed in Germany

Weitere Informationen zu unseren Autoren und Titeln finden Sie unter: www.beltz.de

Inhaltsverzeichnis

Themenblock V: Modellierung der Wirkungen von Unterrichtsqualität (englischsprachig)

Kommentare

Oliver Lüdtke/Alexander Robitzsch

Commentary Regarding the Section "Modelling the Effectiveness of Teaching Quality"

Methodological Challenges in Assessing the Causal Effects of Teaching

Abstract: In this comment paper, we focus on three particular challenges in specifying appropriate models that can be used to estimate the causal effects of teaching in nonrandomized designs. First, we clarify that from a causal perspective the ANCOVA and change score approaches address the same research question (i. e., estimating the causal effects of teaching) but rely on different assumptions to identify the causal effects. Second, we argue that the cumulative effects of teaching (over several years) are often underestimated with two-occasion data. Thereby, we also point out the great potential of marginal structural models for analyzing the effects of time-varying treatments. Finally, we briefly discuss the role of measurement error and compositional effects, which we believe deserve further attention in future methodological research.

Keywords: Causal Effects, ANCOVA, Change Scores, Compositional Effects, Measurement Error

1. Introduction

Assessments of the effects of teaching tend to suffer from several methodological challenges. The articles in this special issue by Naumann, Kuger, Köhler, and Hochweber (in this issue) and Köhler, Kuger, Naumann, and Hartig (in this issue) provide an excellent overview of the many important statistical and methodological developments that have been achieved in the last two decades. In this comment paper, we focus on the particular challenges in specifying appropriate models that can be used to estimate the causal effects of teaching in nonrandomized designs. In line with Naumann et al. (in this issue), we want to show the potential of directed acyclic graphs (DAGs; Pearl, Glymour & Jewell, 2016) for clarifying the – often not articulated – causal assumptions of different modeling choices. More specifically, we use a structural modeling perspective that relies on DAGs to discuss three analytical issues that we believe are particularly relevant in targeting the causal effects of teaching. First, we clarify that the ANCOVA and change score approaches to analyzing two-occasion data discussed by Köhler et al. (in this issue) address the same research question (i. e., estimating the causal effects of teaching) but rely on different assumptions to identify causal effects. Second, we argue that the cumulative effects of teaching (over several years) are often underestimated with two-occasion data (Raudenbush, 2008). Thereby, we introduce a structural model

for three-occasion data and show how the causal effect of a sequence of teaching regimes (e. g., the cumulative effect of teaching across 2 school years) can be estimated. We also point out the great potential of marginal structural models for analyzing the effects of time-varying treatments (Robins, Hernán & Brumback, 2000). Finally, we briefly discuss the role of measurement error and compositional effects, which we believe deserve further attention in future methodological research.

2. ANCOVA versus Change Scores: A Structural Model Perspective

In the following discussion we introduce a structural model for two-occasion data that represents the causal relationships between the variables and allows us to clearly state the causal assumptions that are made by the different analytical approaches (see also Allison, 1990; Kenny, 1975; Kim & Steiner, 2019). More specifically, we assume that a student outcome (e. g., mathematics achievement) is measured at two measurement occasions (e. g., Grades 7 and 8), denoted as Y_1 and Y_2 respectively. We are interested in the effect of a treatment A_2 (e. g., quality of math teaching in Grade 8) on the outcome Y_2. Furthermore, we assume that a confounding variable U (e. g., socioeconomic background, gender) that affects both the student outcomes (Y_1 and Y_2) and the treatment is present. In the interests of simplicity and transparency, we assume that all effects are linear and that the variables are standardized.

To estimate the causal effect of A_2, at least three different approaches can be distinguished (see Köhler et al., in this issue). First, a naive estimator that ignores the pretest measure Y_1 is given by

$$Y_2 = \tau_{\text{naive}} A_2 + \varepsilon \tag{1}$$

Note that the naive estimator is a simple regression of Y_2 on the treatment variable A_2.

Second, an ANCOVA estimator that is conditioned on the pretest measure and has been used in many studies can be represented as

$$Y_2 = \tau_{\text{ANCOVA}} A_2 + \beta_{21} Y_1 + \varepsilon \tag{2}$$

The ANCOVA approach can be considered a special case of a more general class of conditioning methods (e. g., matching methods) in which the causal effect is obtained by conditioning on the pretest (and other observed covariates; see Morgan & Winship, 2015).

Third, a change score approach has been recommended to estimate treatment effects with two-occasion data (e. g., Allison, 1990). In this approach, the difference between the Time 2 and Time 1 scores is regressed on the treatment variable

$$Y_2 - Y_1 = \tau_{\text{change}} A_2 + \varepsilon \tag{3}$$

There has been a longstanding debate among methodologists about whether the ANCOVA approach or the change score approach is more appropriate for analyzing two-occasion data (Lord, 1967). From a descriptive perspective, it can be argued that the two approaches address different questions. In the change score approach, one would be interested in whether differences in the quality of teaching are associated with changes in student achievement. By contrast, the ANCOVA approach estimates whether differences in the quality of teaching predict achievement at Time 2 after controlling for the initial level. However, from a causal perspective, the two approaches address the same question (i. e., estimating the causal effect of the treatment) but rely on different assumptions about potential unobserved confounders. These assumptions can be clarified using the structural model in Figure 1.

It can be shown (see Appendix) that the naive estimator provides an unbiased estimate only if the treatment is unrelated to the pretest and to the unobserved confounder – a condition that is rarely met in nonrandomized designs. The ANCOVA approach produces an unbiased estimate of the treatment effect with two-occasion data if, conditional on Y_1, the unobserved confounder U does not affect the treatment (i. e., $\gamma_A = 0$) or the outcome Y_2 (i. e., $\gamma_2 = 0$). Another view of the ANCOVA approach is that it uses the past outcome and other observed covariates as a proxy for the unobserved confounder (Kim & Steiner, 2019). The change score approach is based on a more subtle set of causal assumptions. First, it is assumed that the treatment is not affected by the past outcome Y_1 (i. e., $\delta = 0$) – an assumption that does not seem very plausible in studies on the effects of teaching. Second, the effect of the unobserved confounder U needs to fulfill a very specific constraint (i. e., $\gamma_2 + \beta\gamma_1 = \gamma_1$) which essentially means that the effects of the (time-invariant) variable U are stable across time (also known as the common trend assumption; Allison, 1990).

Table 1 further illustrates the performances of the ANCOVA and change score approaches under different scenarios. We assumed that the true treatment effect would be

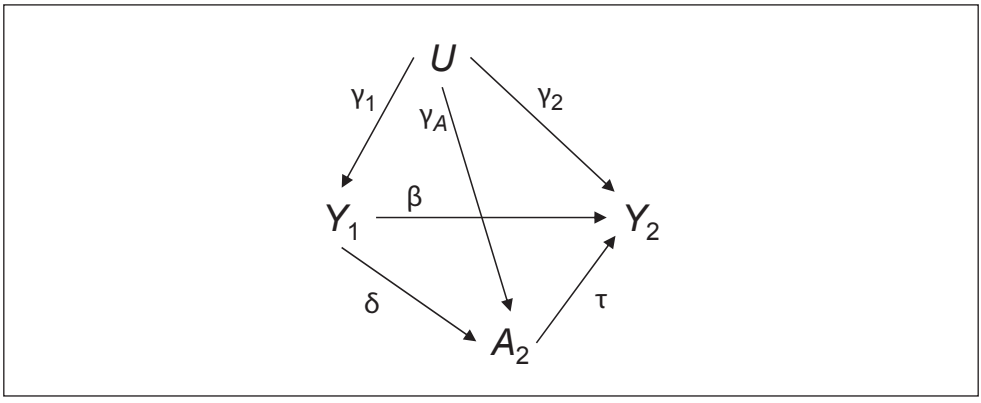

Fig. 1: Structural model for two-occasion data: Effect of treatment A_2 (e. g., quality of teaching) on outcome Y_2 with effects of the past outcome Y_1 and confounder U

γ_2	δ	γ_A	τ_{naive}	τ_{ANCOVA}	τ_{change}
0	0	0	**.20**	**.20**	**.20**
0	0	.3	.26	**.20**	.14
0	.3	0	.35	**.20**	.05
0	.3	.3	.41	**.20**	−.01
.1	0	0	**.20**	**.20**	**.20**
.1	0	.3	.29	.23	.17
.1	.3	0	.36	**.20**	.06
.1	.3	.3	.45	.23	.03
.2	0	0	**.20**	**.20**	**.20**
.2	0	.3	.32	.25	**.20**
.2	.3	0	.37	**.20**	.07
.2	.3	.3	.49	.26	.07

Note. It is assumed that the true treatment effect is $\tau = .20$, the stability of Y is moderate ($\beta = .50$), and the effect of the confounder U on Y_1 is $\gamma_1 = .40$. Unbiased estimates are printed in bold.

Tab. 1: Illustration of bias in the naive, ANCOVA, and change score estimators in two-wave design (see Fig. 1): Size of the estimated treatment effect as a function of γ_2, δ, and γ_A

modest in size (i. e., $\tau = .20$) and that the outcome Y would be moderately stable (i. e., $\beta = .50$), but we manipulated the effect of U on Y_2 (i. e., γ_2) and the treatment A_2 (i. e., γ_A). We also varied whether the past outcome Y_1 had an effect on the treatment (i. e., δ). As expected, the naive estimator in general overestimated the size of the treatment effect, and was only unbiased if the confounder U and the pretest Y_1 were not related to the treatment. The ANCOVA estimator tends to overestimate the true treatment effect and is unbiased under conditions in which U does not have an effect on either Y_2 (i. e., $\gamma_2 = 0$) or the treatment (i. e., $\gamma_A = 0$). By contrast, the change score estimator is only unbiased if $\delta = 0$ and either U does not affect the treatment (i. e., $\gamma_A = 0$) or the common trend assumption is met. Interestingly, the ANCOVA and change score estimators have a useful bracketing property (Angrist & Pischke, 2009; see also Ding & Li, 2019). Under reasonable conditions, the ANCOVA estimator provides an upper bound and the change score provides a lower bound for the true treatment effect.[1]

1 It can be shown that this bracketing property holds under mild assumptions about the data-generating model. More specifically, it needs to be assumed that the (cumulative) effect of U on the outcome is smaller for the posttest than for the pretest – that is, $(1 - \beta)\gamma_1 - \gamma_2 > 0$ – and that $\beta < 1$ (Angrist & Pischke, 2009).

Overall, we tried to clarify that from a causal perspective, the ANCOVA and change score approaches rely on different assumptions for identifying causal effects. Unfortunately, the identifying assumptions of these methods cannot be tested, and in practice it is possible that neither of these assumptions will reflect the true data-generating model. We tend to prefer the ANCOVA approach because it provides a clear rationale for including observed covariates in the analysis (VanderWeele, 2019). However, the change score approach offers the option of controlling for the effect of unobserved confounders. This comes at the price of a very restrictive assumption about the effects of the unobserved confounder (i. e., common trend assumption), an assumption that often does not seem plausible in practice (see Imai & Kim, 2019; Sobel, 2012). In addition, it can be shown that even if the confounder U is observed and included in Equation 3, the change score approach will in general produce biased estimates of the causal effect as long as the past outcome affects the current treatment (see the Appendix).

3. Assessing the Effects of a Sequence of Teaching Experiences

As aptly pointed out by Raudenbush (2008), "whether children can read or reason mathematically is the cumulative result of sequences of teaching experiences over several years" (p. 221). However, the two-occasion design is usually limited to assessing the teaching effects that occur during a single year. To better understand the limitations of two-occasion data for estimating the cumulative effects of teaching, we extended our structural model to include three-occasion data (Fig. 2) in which Y_0 now denotes baseline achievement, and A_1 and A_2 denote a sequence of two treatments (e. g., the quality of teaching over 2 years).

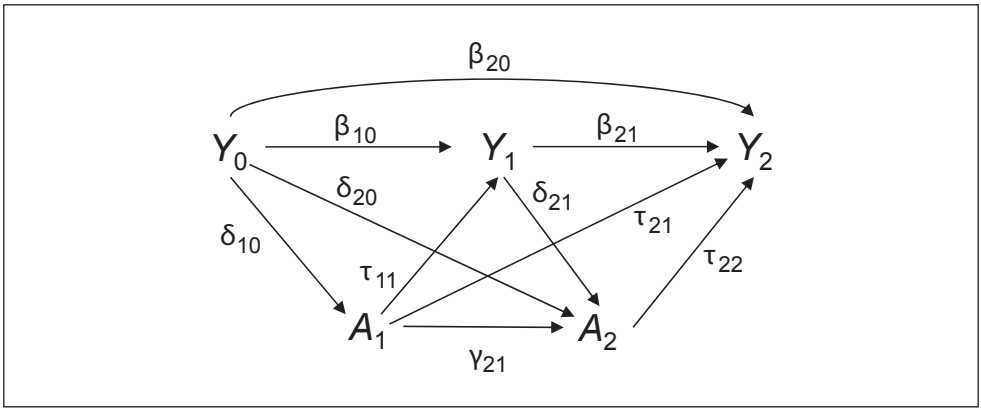

Fig. 2: Structural model for three-occasion data: Effect of the sequence of treatments A_1 and A_2 (e. g., quality of teaching across 2 school years) on outcome Y_2 with the effects of past outcomes Y_0 and Y_1

Estimation of the causal effect of the sequence of treatments A_1 and A_2 (also called an instructional regime; Raudenbush, 2008) on the final student outcome Y_2 thus becomes complicated by the fact that Y_1 is a time-varying confounder that is affected by prior treatment status (i.e., A_1). On the one hand, Y_1 is a confounder of the relationship between A_2 and Y_2. Thus, it is necessary to condition on Y_1 to obtain an unbiased estimate of the effect of A_2. On the other hand, Y_1 is on the causal pathway between past treatment A_1 and Y_2. Thus, controlling for Y_1 would block some of the effect of A_1 on Y_2.

Marginal structural models (MSMs) have been developed as powerful tools that can be used to address issues of time-varying confounders under different potential time-varying treatment regimes (see also Daniel, Cousens, de Stavola, Kenward & Sterne, 2013; Robins et al., 2000). The basic idea of MSMs is to specify a model for the treatment regime in which the effects of confounding variables have been removed. In most cases, MSMs are estimated by weighting methods (Daniel et al., 2013). However, under the assumption of a data-generating model with only linear relations, the coefficients of an MSM can be computed from the coefficients of a path model using tracing rules (Moerkerke, Loeys & Vansteelandt, 2015). In our case, the joint and direct effects of A_1 and A_2 would be given as follows

$$E(Y_2(a_1, a_2)) = (\tau_{21} + \beta_{21}\tau_{11})a_1 + \tau_{22}a_2 \tag{4}$$

where $E(Y_2(a_1, a_2))$ denotes the expected potential outcome that would have resulted if the treatment status for A_1 and A_2 had been set to levels a_1 and a_2, respectively. Thus, the joint effect of increasing the quality of teaching by 1 unit in both time intervals would be $(\tau_{21} + \beta_{21}\tau_{11}) + \tau_{22}$, whereas the direct effects of A_1 and A_2 would be $\tau_{21} + \beta_{21}\tau_{11}$ and τ_{22}, respectively.

In the following, we use the structural model for the three-occasion data (see Figure 2) to illustrate how the ANCOVA or change score approaches that rely on only two-occasion data (i.e., only considering Y_1, Y_2, and A_2) will result in biased estimates of the cumulative effects of a sequence of teaching experiences (see Appendix). We assumed that the outcome Y would show moderate stability across time and that past outcomes (Y_0 and Y_1) would have a small effect on the current treatment (A_1 and A_2). In Table 2, we varied the effects of A_1 (i.e., τ_{11} and τ_{21}) and A_2 (i.e., τ_{22}) and the stability of the treatment (i.e., γ_{21}). The MSM estimates (see Equation 4) provide the joint effect of the treatment sequence. For example, in the penultimate row, the joint effect of increasing both treatments by 1 unit is given by $(.1 + .55 \cdot .2) + .2 = .41$, which is the sum of the direct effect of A_1 and the direct effect of A_2. However, both the ANCOVA and change score approaches underestimate the joint effect of the treatment. Note that the ANCOVA approach is particularly biased when the treatment shows only moderate stability (i.e., $\gamma_{21} \leq .4$). This is a reasonable scenario when classes change their teacher after a year. In practice, effects of teaching (or observed covariates) are expected to deviate from linearity (see Naumann et al., in this issue). In this case, the MSMs would also be nonlinear, and weighting or Monte Carlo-based approaches would be recommended (Daniel

τ_{11}	τ_{21}	τ_{22}	MSM			$\gamma_{21} = 0$		$\gamma_{21} = .4$		$\gamma_{21} = .8$	
			Joint	A_1	A_2	τ_{ANCOVA}	τ_{change}	τ_{ANCOVA}	τ_{change}	τ_{ANCOVA}	τ_{change}
0	0	0	.00	.00	.00	.02	−.05	.05	−.05	.08	−.04
0	.1	0	.10	.10	.00	.02	−.04	.09	.00	.17	.04
0	.2	0	.20	.20	.00	.02	−.04	.14	.05	.27	.13
.1	0	.1	.16	.06	.10	.12	.05	.14	.04	.16	.02
.1	.1	.1	.26	.16	.10	.12	.06	.18	.08	.26	.11
.1	.2	.1	.36	.26	.10	.12	.07	.23	.13	.36	.20
.2	0	.2	.31	.11	.20	.22	.15	.23	.12	.24	.09
.2	.1	.2	.41	.21	.20	.22	.16	.27	.17	.34	.18
.2	.2	.2	.51	.31	.20	.22	.17	.31	.22	.45	.27

Note. MSM = Marginal structural model. It is assumed that the outcome *Y* shows moderate stability across time ($\beta_{10} = .60$, $\beta_{20} = .30$, and $\beta_{21} = .55$) and that past outcomes affect the current treatment ($\delta_{10} = .30$, $\delta_{20} = .10$, and $\delta_{21} = .20$). MSM estimates are based on Equation 4.

Tab. 2: Illustration of bias in the ANCOVA and change score estimators with the three-occasion data (see Fig. 2) as a function of the effects of A_1 and A_2 and the stability of the treatment (γ_{21})

et al., 2013). We believe that MSMs have great potential and deserve more attention in research on the effectiveness of learning and teaching (see Vandecandelaere, Vansteelandt, De Fraine & Van Damme, 2016).

4. Further Challenges in Estimation of the Causal Effects of Teaching

In our discussion of different approaches for estimating the causal effects of teaching, we have made several simplifying assumptions. First, we did not mention the multilevel structure of educational data. Usually, multilevel models are applied to take into account a nested data structure, and to estimate the effects of variables that are located at different levels. It should be emphasized that our remarks about the performance of the ANCOVA or change score estimators would also apply to specification of the structural model at the class level in multilevel structural equation models (MSEMs). As pointed out by Naumann et al. (in this issue), measures of teaching are affected by different kinds of error (e. g., sampling error, measurement error; Kane & Brennan, 1977). MSEMs provide a powerful tool that can be used to take these errors into account when

estimating the effects of teaching, but could provide unstable estimates in certain data constellations (e. g., a small number of classes, many items, low intraclass correlations). Bayesian methods have been shown to provide improved parameter estimates even under such challenging conditions (Zitzmann, Lüdtke, Robitzsch & Marsh, 2016). Alternatively, estimation of the measurement model (i. e., model for items) could be separated from estimation of the structural model, which could also result in more stable and robust estimates (see Anderson & Gerbing, 1982).[2]

Second, the correct way to treat compositional effects can be debated. More specifically, when conditioning on the pretest measure (e. g., Y_1 in Fig. 1 and 2), it is a crucial question as to whether the class mean (or school mean) should also be included in the regression. MSEMs decompose Level 1 predictors into a within-part and a between-part, and the group means of the Level 1 predictors are introduced into the model by default (Rabe-Hesketh, Skrondal & Zheng, 2012). This strategy of including the group means and controlling for compositional effects was also recommended by Köhler et al. (in this issue), who noted that "the modeling of T1 at L2 is vital, because otherwise the previous average class level would not be controlled for" (p. 204). However, it has been argued that controlling for compositional effects can bias the potential effects of teaching quality (Castellano, Rabe-Hesketh & Skrondal, 2014). Imagine that in the transition from elementary to secondary school, students with more favorable background characteristics are more likely to be sent to better schools (e. g., schools with greater resources, more motivated staff, better expected performance). Furthermore, it could be possible that better teachers (i. e., higher teaching quality) are attracted by better schools. As can be seen in the structural model in Figure 3, this would result in positive associations of student achievement Y_1 (more exactly, its between-part Y_{B1}) as well as the treatment A_2 with the random school effect U on the posttest (i. e., $\rho_{Y_{B1}U}\sigma_U > 0$, and $\rho_{A_2U}\sigma_U > 0$). Note that the pretest is determined before U and affects the grouping of students into different schools, resulting in an artificially increased composition effect (Castellano, Rabe-Hesketh & Skrondal, 2014; see also Cronbach, 1976). Hence, the positive covariance $\rho_{Y_{B1}U}\sigma_U$ will positively bias the estimate of β_B (i. e., "overcontrolling" for compositional effects; see the Appendix), which in turn could negatively bias the estimate of the treatment effect τ. However, the ANCOVA estimator could also be positively biased if higher teaching quality is associated with better schools. In practice, it is likely that both bias contributions are present and the ANCOVA estimator would be unbiased in the special case that they cancel each other out (i. e., $\rho_{A_2U} - \rho_{Y_{B1}U}\rho_{Y_{B1}A_2} = 0$). Interestingly, the change score estimator has the potential to control for the artificial grouping effect of students (unlike the ANCOVA), but will still be biased if the treatment is affected by the pretest, and if the correlation between the pretest and posttest differs sub-

2 For example, in generalizability theory, less parameterized measurement models are used to decompose the different error components (Brennan, 2001). Further integration of these measurement models would be a promising way to obtain more stable estimates in multilevel models.

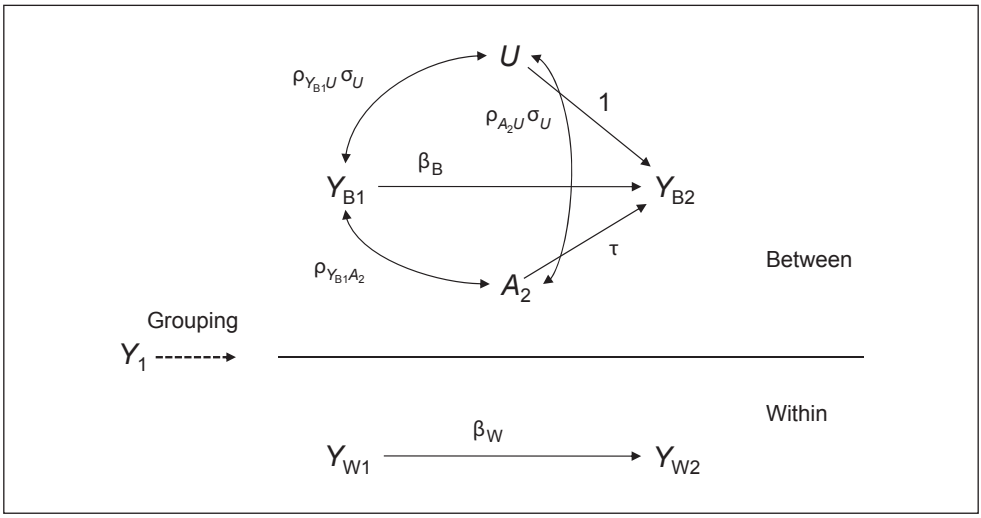

Fig. 3: Structural model for the role of composition effects with two-occasion data. The assignment to different clusters depends on the pretest Y_1 (e.g., children with high pretest scores are more likely to be sent to better schools by their parents), resulting in a covariance between Y_{B1} and U ($\rho_{Y_{B1}U}\sigma_U$).

stantially from one. Again, this illustrates how a structural model perspective can help to clarify the assumptions behind different modeling approaches.

References

Allison, P. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology, 20,* 93–114.

Anderson, J. W., & Gerbing, D. W. (1982). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, context, and endogeneity in school and teacher comparisons. *Journal of Educational and Behavioral Statistics, 39,* 333–367.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium.

Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., & Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. *Statistics in Medicine, 32,* 1584–1618.

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis.* https://doi.org/10.1017/pan.2019.25.

Imai, K., & Kim, I. S. (2019). When should we use fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science, 63,* 467–490.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47,* 267–292.
Kenny, D. A. (1975). A quasi-experimental approach to assessing treatment effects in the non-equivalent control group design. *Psychological Bulletin, 82,* 345–362.
Kim, Y., & Steiner, P. M. (2019). Gain scores revisited: A graphical modeling perspective. *Sociological Methods and Research.* https://doi.org/10.1177/0049124119826155.
Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.
Moerkerke, B., Loeys, T., & Vansteelandt, S. (2015). Structural equation modeling versus marginal structural modeling for assessing mediation in the presence of posttreatment confounding. *Psychological Methods, 20,* 204–220.
Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). Cambridge: University Press.
Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer.* John Wiley & Sons.
Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2012). Multilevel structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 512–531). New York, NY: Guilford.
Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45,* 206–230.
Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology, 11,* 550–560.
Sobel, M. E. (2012). Does marriage boost men's wages? Identification of treatment effects in fixed effects regression models for panel data. *Journal of the American Statistical Association, 107,* 521–529.
Vandecandelaere, M., Vansteelandt, S., De Fraine, B., & Van Damme, J. (2016). Time-varying treatments in observational studies: Marginal structural models of the effects of early grade retention on math achievement. *Multivariate Behavioral Research, 51,* 843–864.
VanderWeele, T. (2019). Principles of confounder selection. *European Journal of Epidemiology, 34,* 211–219.
Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling, 23,* 661–679.

Appendix: Derivations of Bias for the ANCOVA and Change Score Approaches

In this Appendix, we sketch the bias derivations for the ANCOVA and change score approaches. All variables are assumed to be standardized.

Two-Occasion Data

Given the structural model in Figure 1, the covariances between the observed variables Y_1, A_2, and Y_2 are derived as follows: $\text{Cov}(A_2, Y_1) = \delta + \gamma_1\gamma_A$, $\text{Cov}(Y_2, Y_1) = \beta + \delta\tau + \gamma_1\gamma_2 + \gamma_1\gamma_A\tau$, $\text{Cov}(Y_2, A_2) = \tau + \beta\delta + \gamma_2\gamma_A + \beta\gamma_1\gamma_A + \delta\gamma_1\gamma_2$. Note that the naive estimator τ_{naive} is given by $\text{Cov}(Y_2, A_2)$. The ANCOVA estimator (without controlling for the unobserved confounder U) is given as follows:

$$\tau_{ANCOVA} = \frac{Cov(Y_2, A_2) - Cov(Y_2, Y_1)\, Cov(A_2, Y_1)}{1 - Cov(A_2, Y_1)^2} = \tau + \frac{\gamma_2 \gamma_A (1 - \gamma_1^2)}{1 - 2\delta\gamma_1\gamma_A - \delta^2 - \gamma_1^2\gamma_A^2} \quad (A.1)$$

The ANCOVA estimator is unbiased if $\gamma_A = 0$ or if $\gamma_2 = 0$. Furthermore, the change score estimator is given by

$$\tau_{change} = Cov(Y_2, A_2) - Cov(Y_1, A_2) = \tau + \gamma_A(\gamma_2 + \beta\gamma_1 - \gamma_1) + \delta(\beta + \gamma_1\gamma_2 - 1) \quad (A.2)$$

The change score estimator is unbiased if $\delta = 0$ and if $\gamma_2 + \beta\gamma_1 = \gamma_1$ (i. e., the effect of U is stable with respect to Y_1 and Y_2). In addition, it is evident that the ANCOVA estimator is unbiased if the confounder U is included in the regression in Equation 2. However, it can be shown that the change score estimator is not unbiased even if the confounder U is included in the regression in Equation 3

$$\tau_{change,U} = \frac{Cov(Y_2 - Y_1, A_2) - Cov(Y_2 - Y_1, U)\, Cov(A_2, U)}{1 - Cov(A_2, U)^2} = \tau + \frac{\delta(1 - \beta)\,(1 - \gamma_1^2)}{1 - 2\delta\gamma_1\gamma_A - \gamma_A^2 - \delta^2\gamma_1^2}$$

In this case, the change score estimator would be unbiased if $\delta = 0$ (i. e., past outcome Y_1 does not affect the treatment).

Three-Occasion Data

The structural model for the three-occasion data consists of five observed variables (i. e., Y_0, Y_1, Y_2, A_1, and A_2). The implied covariances between Y_1, Y_2, and A_2 are given as follows:

$$Cov(A_2, Y_1) = \delta_{21} + \beta_{10}\,\delta_{20} + \gamma_{21}\,\tau_{11} + \beta_{10}\,\delta_{10}\,\gamma_{21} + \delta_{10}\,\delta_{20}\,\tau_{11}$$

$$\begin{aligned} Cov(Y_2, Y_1) = {} & \beta_{21} + \beta_{10}\,\beta_{20} + \delta_{21}\,\tau_{22} + \tau_{11}\,\tau_{21} + \beta_{10}\,\delta_{10}\,\tau_{21} + \beta_{10}\,\delta_{20}\,\tau_{22} + \beta_{21} + \beta_{10}\,\beta_{20} \\ & + \delta_{21}\,\tau_{22} + \tau_{11}\,\tau_{21} + \beta_{10}\,\delta_{10}\,\tau_{21} + \beta_{10}\,\delta_{20}\,\tau_{22} \end{aligned}$$

$$\begin{aligned} Cov(Y_2, A_2) = {} & \tau_{22} + \beta_{20}\delta_{20} + \beta_{21}\,\delta_{21} + \gamma_{21}\,\tau_{21} + \beta_{10}\,\beta_{20}\,\delta_{21} + \beta_{10}\,\beta_{21}\,\delta_{20} + \beta_{20}\,\delta_{10}\,\gamma_{21} + \\ & \beta_{21}\,\gamma_{21}\,\tau_{11} + \delta_{10}\,\delta_{20}\,\tau_{21} + \delta_{21}\,\tau_{11}\,\tau_{21} + \beta_{10}\,\beta_{21}\,\delta_{10}\,\gamma_{21} + \beta_{10}\,\delta_{10}\,\delta_{21}\,\tau_{21} + \\ & \beta_{20}\,\delta_{10}\,\delta_{21}\,\tau_{11} + \beta_{21}\,\delta_{10}\,\delta_{20}\,\tau_{11} \end{aligned}$$

The ANCOVA and change score estimators that ignore the baseline measure Y_0 and the previous treatment A_1 can now be derived by inserting the covariances into Equations A.1 and A.2. The calculations, however, are cumbersome and do not provide any further insights.

Role of Compositional Effects

The covariances between the observed variables at the between level are given as follows (see Fig. 3): $\mathrm{Cov}(A_2, Y_{B1}) = \rho_{Y_{B1}A_2}$; $\mathrm{Cov}(Y_{B2}, Y_{B1}) = \beta_B + \tau\rho_{Y_{B1}A_2} + \rho_{Y_{B1}U}\sigma_U$; $\mathrm{Cov}(Y_{B2}, A_2) = \tau + \beta_B\rho_{Y_{B1}A_2} + \rho_{A_2U}\sigma_U$. The between group coefficient of the pretest Y_{B1} is given as

$$\hat{\beta}_B = \beta_B + \frac{(\rho_{Y_{B1}U} - \rho_{A_2U}\rho_{Y_{B1}A_2})\sigma_U}{1 - \rho^2_{Y_{B1}A_2}}$$

which is positively biased if $\rho_{Y_{B1}U} - \rho_{Y_{B1}A_2}\rho_{A_2U} > 0$. Typically, the ANCOVA estimator of the treatment effect at the between level is biased

$$\tau_{\mathrm{ANCOVA}} = \tau + \frac{(\rho_{A_2U} - \rho_{Y_{B1}U}\rho_{Y_{B1}A_2})\sigma_U}{1 - \rho^2_{Y_{B1}A_2}}$$

It overadjusts for the compositional effect if $\rho_{A_2U} - \rho_{Y_{B1}U}\rho_{Y_{B1}A_2} < 0$ which, in turn, results in a negatively biased treatment effect estimate. The change score estimator can be calculated as

$$\tau_{\mathrm{change}} = \tau + \rho_{A_2U}\sigma_U - (1 - \beta_B)\rho_{Y_{B1}A_2}$$

If $1 - \rho^2_{Y_{B1}A_2} \approx 1$, it can be seen that the change score estimator has a lower potential for negative bias if $1 - \beta_B < \rho_{Y_{B1}U}\sigma_U$, which is fulfilled if pretest Y_{B1} and posttest Y_{B2} are highly correlated at the between level.

Zusammenfassung: Der vorliegende Kommentar konzentriert sich auf drei Herausforderungen, die bei der Spezifikation des Analysemodells auftreten. Erstens wird gezeigt, welche Annahmen über die Wirkung konfundierender Variablen sowohl mit dem ANCOVA- als auch mit dem Differenzwert-Ansatz getroffen werden müssen. Zweitens wird argumentiert, dass die kumulativen Effekte des Unterrichts (über mehrere Jahre) mit Zwei-Wellen-Daten häufig unterschätzt werden. Dabei wird das große analytische Potential von Marginal Structural Models betont, die sich besonders zur Schätzung zeitlich variierender kausaler Effekte eignen. Abschließend werden mit der Rolle des Messfehlers und der Behandlung von Kompositionseffekten zwei Themen diskutiert, die aus unserer Sicht in zukünftiger Forschung noch mehr beachtet werden sollten.

Schlagworte: kausale Effekte, ANCOVA, Differenzwerte, Kompositionseffekte, Messfehler

Contact

Prof. Dr. Oliver Lüdtke, IPN – Leibniz Institute for Science and Mathematics Education,
Department of Educational Measurement,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: oluedtke@leibniz-ipn.de

Dr. Alexander Robitzsch, IPN – Leibniz Institute for Science and Mathematics Education,
Department of Educational Measurement,
Olshausenstr. 62, 24118 Kiel, Germany
E-Mail: robitzsch@leibniz-ipn.de