Behavior Research Methods (2018) 50:1581–1601
https://doi.org/10.3758/s13428-018-1031-x

Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS

Wolfgang Wiedermann[1] · Xintong Li[1]

Published online: 16 April 2018

Abstract

In nonexperimental data, at least three possible explanations exist for the association of two variables x and y: (1) x is the cause of y, (2) y is the cause of x, or (3) an unmeasured confounder is present. Statistical tests that identify which of the three explanatory models fits best would be a useful adjunct to the use of theory alone. The present article introduces one such statistical method, direction dependence analysis (DDA), which assesses the relative plausibility of the three explanatory models on the basis of higher-moment information about the variables (i.e., skewness and kurtosis). DDA involves the evaluation of three properties of the data: (1) the observed distributions of the variables, (2) the residual distributions of the competing models, and (3) the independence properties of the predictors and residuals of the competing models. When the observed variables are nonnormally distributed, we show that DDA components can be used to uniquely identify each explanatory model. Statistical inference methods for model selection are presented, and macros to implement DDA in SPSS are provided. An empirical example is given to illustrate the approach. Conceptual and empirical considerations are discussed for best-practice applications in psychological data, and sample size recommendations based on previous simulation studies are provided.

Keywords Linear regression model · Direction of effects · Direction dependence · Observational data · Nonnormality

This article introduces methods of direction dependence and presents a unified statistical framework to discern the causal direction of effects in linear models using observational data. Existing regression-type methods allow researchers to quantify the magnitude of hypothesized effects but are of limited use when establishing the direction of effects between variables—that is, whether $x \rightarrow y$ or $y \rightarrow x$ correctly describes the causal flow between two variables x and y. The statistical framework proposed in this article allows researchers to make conclusions about the direction of effects. In the present work, we focus on observational (nonexperimental) data settings because this type of data is the chief material for the presented principles of direction dependence. The issue of effect directionality in the context of experimental studies (e.g., when decomposing total effects into direct and indirect effect components) will be taken up in the Discussion section.

Establishing cause–effect relations between variables is a central aim of many empirical studies in the social sciences. The direction of influence between variables is a key element of any causal theory that purports to explain the data-generating mechanism (Bollen, 1989). Questions concerning direction of effect naturally arise in observational studies. For example, it may not be entirely clear whether tobacco consumption causes depression and anxiety or whether people with symptoms of depression and anxiety are more likely to engage in health damaging behavior (Munafò & Araya, 2010; Taylor et al., 2014); whether violent video games expose players to aggressive behavior or whether aggressive people are simply more attracted to violent video games (Gentile, Lynch, Linder, & Walsh, 2004); or whether lead exposure contributes to the development of ADHD or whether children with ADHD symptoms are unable to stay focused enough to avoid lead-tainted objects (Nigg et al., 2008).

Unfortunately, also useful in assessing the magnitude and statistical significance of an assumed causal effect, standard regression-based methods are of limited use when addressing directionality issues per se. When an association between x

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13428-018-1031-x) contains supplementary material, which is available to authorized users.

✉ Wolfgang Wiedermann
wiedermannw@missouri.edu

[1] Statistics, Measurement, and Evaluation in Education, Department of Educational, School, and Counseling Psychology, College of Education, University of Missouri, 13B Hill Hall, Columbia, MO 65211, USA

and y exists, at least three possible explanations can be entertained: (1) x causes y ($x \rightarrow y$), (2) y causes x ($y \rightarrow x$), and (3) neither relation exists due to a spurious association of both variables with a third variable (sometimes termed a "confounder"; see Fig. 1).[1] The Pearson product-moment correlation and ordinary least square (OLS) estimates do not adjudicate regarding the model that best represents the data-generating mechanism (von Eye & DeShon, 2012). Researchers who use regression models must therefore make their decision as to the direction of effect on the basis of a priori theory and substantive arguments. However, statistical tools often are desirable to empirically demonstrate the explanatory superiority of one theory over plausible alternatives. The present contribution introduces such a tool—direction dependence analysis (DDA; Wiedermann & von Eye, 2015a). Although standard regression models use only estimates of first- and second-order moments (i.e., means, variances, and covariances) to assess the magnitude and statistical significance of regression weights, DDA, by contrast, uses estimates of higher-order moments (i.e., skewness and kurtosis) to assess the relative plausibility of directional alternatives.

Methods of causal discovery have experienced rapid development within the last decades and various causal search algorithms have been proposed (see Spirtes & Zhang, 2016, for a recent overview). These search algorithms are designed to learn plausible causal structures from multivariate data. Examples of such algorithms include the PC algorithm (Spirtes, Glymour, & Scheines, 2000), greedy equivalence search (Chickering, 2002), cyclic causal discovery (Richardson & Spirtes, 1999), fast causal inference (Zhang, 2008), and linear non-Gaussian acyclic models (cf. Shimizu, 2016; Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006) that are either designed to discover (Markov) equivalence classes of directed acyclic graphs (DAGs; i.e., a small subset of candidate models that have the same support by the data in terms of model fit; cf. Verma & Pearl, 1991) or uncover DAG structures beyond equivalence classes. All these algorithms constitute important exploratory tools for causal learning and are, thus, ideally suited to generate new substantive hypotheses concerning the causal nature of constructs.

DDA, in contrast, is concerned with a research scenario that is confirmatory in nature—that is, situations in which a substantive theory about the causal relation exists and the researcher wishes to know if the causal direction assumed by the model is plausible relative to the alternative scenarios a reasonable skeptic might propose. The primary goal is to probe this causal theory against alternatives while adjusting for potential background variables known to be explanatory in nature. Thus, instead of extracting plausible DAG structures (or classes of equivalent DAGs) for a given dataset, one is interested in testing a specific model (e.g., lead exposure → ADHD) against a plausible alternative (ADHD → lead exposure). The present article is designed to introduce principles of DDA to quantitative researchers. Although previous work (e.g., Dodge & Rousson, 2000, 2001; Dodge & Yadegari, 2010; Muddapur, 2003; Sungur, 2005; Wiedermann & von Eye, 2015a, 2015b, 2015c) has focused on direction dependence methods to choose between the two models $x \rightarrow y$ and $y \rightarrow x$ (Figs. 1a and b), direction dependency in the presence of confounders has received considerably less attention. To fill this void, we present extensions of DDA to scenarios in which confounders are present and incorporate these new insights into the existing direction dependence principle. As a result, we propose a unified framework that allows one to identify each explanatory model given in Fig. 1.

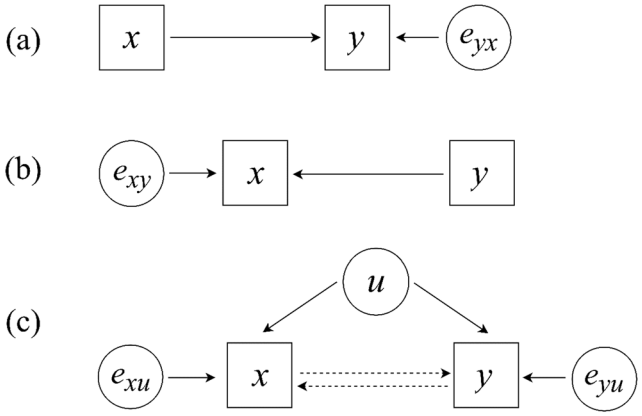

Fig. 1 Conceptual diagrams of three possible models to explain the association between two variables x and y. Squares are used for observed variables and circles are used for latent variables (with error terms denoted with e). Model a = directional effect of x on y; Model b = directional effect of y on x; Model c = confounded x–y association due to an unmeasured third variable u. The two dashed arrows in Model c indicate that either $x \rightarrow y$ or $y \rightarrow x$ can be biased by confounding

The article is structured as follows: First, we define statistical models suitable for DDA and introduce model assumptions. We then introduce elements of DDA and summarize new results that describe the behavior of DDA tests when unmeasured confounders are present. Next, three SPSS macros are introduced that make DDA accessible to applied researchers and a data example is given to illustrate their application. The article closes with a discussion of conceptual and empirical requirements of DDA, potential data-analytic pitfalls, and potential extensions of the DDA methodology. In addition, sample size recommendations based on previous simulation studies are provided.

[1] Reciprocal causal models (x affects y, and vice versa) may serve as a fourth possible explanation for variable associations. Although it is mathematically possible to estimate reciprocal effects with cross-sectional data (James & Singh, 1978), some controversy exists about the adequacy of those estimates (Wong & Law, 1999) due to the absence of temporality. In some theories, temporal precedence constitutes a crucial element to quantify feedback loops, and thus, longitudinal data are usually preferred (Rogosa, 1985).

The direction dependence principle

Direction dependence can be defined as the asymmetry between cause and effect. The model $x \rightarrow y$ implies that changing x (the cause) changes y (the effect) but changing y does not lead to change in x (see also Pearl, 2009; Peters, Janzing, & Schölkopf, 2017). Reversely, when changing y changes x but, at the same time, changing x does not change y, then the model $y \rightarrow x$ describes the causal relation. Limitations of conventional association-based approaches to uncover the asymmetry of cause and effect can be explained by the fact that these methods only consider variation up to the second order moments and, thus, rely on correlation in its symmetric form—that is, $\text{cor}(x, y) = \text{cor}(y, x)$. The key element of DDA is to consider variable information beyond second order moments (specifically skewness and kurtosis) because asymmetry properties of the Pearson correlation and the related linear model appear under nonnormality. These asymmetry properties are of importance when x and y are not exchangeable in their roles as explanatory and response variables without leading to systematic model violations. DDA, thus, requires and makes use of nonnormality of variables to gain deeper insight into the causal mechanism. DDA consists of three core components: (1) distributional properties of observed variables, (2) distributional properties of error terms of competing models, and (3) independence properties of error terms and predictors in the competing models. On the basis of new results regarding direction dependence in the presence of confounders, we show that unique patterns of DDA component outcomes exist for each of the three models shown in Fig. 1. These outcome patterns enable researchers to select between competing models. In the following paragraphs, we formally define the statistical models considered. We then introduce DDA components separately for confounder-free "true" models and when confounders are present. In addition, statistical inference compatible with direction dependence is discussed. To simplify presentation, we assume that $x \rightarrow y$ corresponds to the "true" model and $y \rightarrow x$ represents the directionally mis-specified model.

Model definitions

We start the introduction to DDA by defining the statistical models considered. Although statistical models can either be used for the purposes of *explanation* or *prediction*[2] (Geisser, 1993), DDA is designed for the task of validating explanatory models, that is, to test the causal hypothesis assumed under a given theory. Assume that a construct $\mathcal{X}$ (e.g., lead exposure) causes construct $\mathcal{Y}$ (e.g., ADHD symptomatology) through mechanism $\mathcal{F}$ —that is, $\mathcal{Y} = \mathcal{F}(\mathcal{X})$. Furthermore, let x and y be operationalizations of $\mathcal{X}$ and $\mathcal{Y}$ (e.g., blood lead concentration and number of DSM-IV hyperactive–impulsive symptoms) and define f as the statistical model (e.g., the linear model) to approximate $\mathcal{F}$ —that is, $y = f(x)$. The direction dependence framework provides a set of statistical tools to evaluate the directionality assumption of $y = f(x)$ implied by the causal theory $\mathcal{X} \rightarrow \mathcal{Y}$.

DDA assumes that $\mathcal{X}$ is a *nonnormally distributed* construct whose cause lies outside the causal mechanism $\mathcal{X} \rightarrow \mathcal{Y}$ and that the two observed variables, x and y, are continuous. Furthermore, we assume that the data-generating mechanism is recursive in nature (i.e., the causal flow is unidirectional without feedback loops) and can be approximated by the linear model—that is, given that $\mathcal{X} \rightarrow \mathcal{Y}$ constitutes the "true" mechanism, $y = f(x)$ can be written as (cf. Fig. 1a; for simplicity, but without loss of generality, we assume that the intercept is fixed at zero)

$$y = b_{yx}x + e_{yx} \tag{1}$$

The slope b_{yx} denotes the change in the fitted value of y for a one-unit increase in x and represents the causal effect of $x \rightarrow y$. Estimates of the causal effect are usually obtained using OLS or, in structural equation models (SEMs), maximum likelihood estimation. Nonnormality of the "true" predictor, quantified as nonzero skewness $\gamma_x = E[(x-E[x])^3]/\sigma_x^3$ and/or nonzero excess kurtosis $\kappa_x = E[(x-E[x])^4]/\sigma_x^4-3$ (with E being the expected value operator), is assumed to reflect an inherent distributional characteristic of $\mathcal{X}$ (as opposed to nonnormality due to boundaries of the operationalized x). The error term e_{yx} is assumed to be normally distributed (with zero mean and variance $\sigma^2_{e_{yx}}$), serially independent, and independent of x.

When $\mathcal{Y} \rightarrow \mathcal{X}$ instead of $\mathcal{X} \rightarrow \mathcal{Y}$ describes the causal mechanism (Fig. 1b), the corresponding linear model is

$$x = b_{xy}y + e_{xy} \tag{2}$$

where b_{xy} captures the causal effect of y on x. Here, y represents a nonnormal external influence and e_{xy} denotes a normally distributed error term with zero mean and variance $\sigma^2_{e_{xy}}$, which is serially independent and independent of y.

Finally, as a third possible explanation, we consider an unknown confounding construct $\mathcal{U}$ (and its continuously operationalized variable u) that has a causal effect on both $\mathcal{X}$ and $\mathcal{Y}$. A conceptual diagram of a confounded x–y relation is given in Fig. 1c. The dashed arrows in Fig. 1c indicate that either the model $x \rightarrow y$ in Eq. 1 or $y \rightarrow x$ in Eq. 2 can be biased by confounding. For the purpose of introducing the confounder

[2] Although both types of statistical models are of importance for theory building (Braun & Oswald, 2011) and can be characterized as association-based models in the context of observational data, each one plays a different role in the process of testing and redefining theories. In explanatory models, a priori theories carry the crucial element of causation and the goal is to match a statistical model f and an underlying mechanism $\mathcal{F}$ and use x and y as tools to estimate and validate f for the purpose of testing the causal hypothesis of interest. In contrast, in predictive models, f is considered being a tool capturing variable associations and x and y are of primary interest to build valid models for the purpose of forecasting new response values (Shmueli, 2010).

model, we focus on the model $x \rightarrow y$. In this case, the model in Eq. 1 changes to

$$\begin{aligned} x &= b_{xu}u + e_{xu} \\ y &= b_{yx}x + b_{yu}u + e_{yu} \end{aligned} \tag{3}$$

where b_{xu} and b_{yu} denote the effects of the confounder on x and y, respectively. The considered setup also includes the case of "pure" confounding as a special case (i.e., the observed association between x and y is entirely attributable to the existence of u) when $b_{yx} = 0$. In Model 3, u is assumed to be a nonnormally distributed external influence, and e_{xu} and e_{yu} are normally distributed error terms (exhibiting zero means and variances $\sigma^2_{e_{xu}}$ and $\sigma^2_{e_{yu}}$) that are independent of u and of each other.

It is common to include covariates (e.g., background or baseline measures) in statistical models to increase precision of parameter estimates and statistical power. In other words, covariates are independent variables that are considered in a target model to control for their influences on the putative response. In contrast, failing to include important covariates can lead to confounded parameter estimates when covariates are (in addition to their relation to the response) correlated with other predictors. However, several authors have cautioned against careless use of covariates because conditioning on covariates can also increase the bias of causal estimates (Pearl, 2009; Spirtes, Richardson, Meek, Scheines, & Glymour, 1998). Similar considerations hold for statistical models in the context of DDA. To be eligible for DDA, covariates must be known to be on the explanatory side of the statistical model. In addition, one must ensure that a recursive causal ordering of the covariates themselves is theoretically possible and that all covariates can be expressed as linear combinations of mutually independent external influences. We can formally express these prerequisites for a given set of covariates z_j (j = 1, ... , J) as $z_j = \sum_{k(i) \leq k(j)} a_{ji}\eta_i$, with $k(i) \leq k(j)$ describing the causal order of covariates (i.e., z_i precedes z_j). The parameter a_{ji} describes the total effect and η_i denoting the external influence associated with z_i. When no other covariate precedes z_j, one obtains $z_j = \eta_j$ with $a_{ji} = 1$. For example, suppose that two covariates (z_1 and z_2) are known to influence y and that z_1 precedes z_2 and no other covariate precedes z_1 (i.e., $z_1 = \eta_1$). In this case, one obtains $z_2 = a_{21}\eta_1$ with $z_1 = \eta_1$, which implies that z_2 can be expressed as a (weighted) external influence. Consider again the example of ADHD and blood lead exposure. Two factors that are known to affect ADHD symptomology are prenatal maternal emotional stress (Harris & Seckl, 2011) and the cultural context of the child (Miller, Nigg, & Miller, 2009; see also Nigg, 2012). Arguments of temporality or logical order of effects can be used to evaluate the eligibility of covariates for DDA. Both, prenatal maternal stress and cultural context are located earlier in time than the child's blood lead level and ADHD symptomology under study that justifies their use as background variables. Furthermore, in principle, we are also able to establish a causal order of the covariates themselves—that is, cultural context may be conceived as a background variable directly or indirectly contributing to maternal stress level. In other words, both variables are unlikely to render the target model cyclic, which makes them eligible to be covariates in DDA. In general, covariates can be either continuous or categorical. For categorical covariates, however, we need to assume that these variables constitute external influences themselves—that is, we exclude cases in which categorical variables serve as outcomes of other independent variables in the model (detailed explanations will be given below). Although this assumption is stricter than the continuous case, it still allows multiple-group scenarios in which the magnitude of the causal effect of predictor and outcome can vary across groups. When categorical covariates are present, a two-stage approach of model estimation is preferable. That is, in a first step, the effect of categorical covariates is partialled out of the putative predictor (e.g., x), the putative outcome (y), and all the continuous covariates and extracted regression residuals from these auxiliary models are subsequently used as "purified" measures (an example is given below). According to the Frisch–Waugh–Lovell theorem (cf. Frisch & Waugh, 1933; Lovell, 1963; sometimes called the *regression anatomy formula*: Angrist & Pischke, 2009), regressing the "purified" outcome on the "purified" independent variables in the second step leads to the same model parameters as in the full multiple regression model including categorical covariates.

DDA Component I: Distributional properties of observed variables

Absence of confounders Asymmetry properties in terms of observed variable distributions emerge from the additive nature of the linear model—that is, a response is defined as the sum of a (nonnormally distributed) explanatory part and a (normally distributed) error component. Intuitively, distributional differences of predictor and response variables emerge because the response is defined as the convolution of a nonnormal and a normal variate. In other words, adding a normal error term to a nonnormal predictor will necessarily cause the response to be more normally distributed than the predictor. Dodge and Rousson (2000, 2001) as well as Dodge and Yadegari (2010) presented algebraic proofs for this relation and showed that the Pearson correlation ρ_{xy} has asymmetric properties when considering higher moments of x and y.

Specifically, the cube of ρ_{xy} can be expressed as the ratio of skewness of response and predictor,[3]

$$\rho_{xy}^3 = \frac{\gamma_y}{\gamma_x} \tag{4}$$

(as long as $\gamma_x \neq 0$) and the fourth power of ρ_{xy} can be written as the ratio of excess kurtosis of response and predictor,

$$\rho_{xy}^4 = \frac{\kappa_y}{\kappa_x} \tag{5}$$

(as long as $\kappa_x \neq 0$). Because ρ_{xy} is bounded on the interval [–1, 1], absolute values of skewness and excess kurtosis of the response y will always be smaller than absolute skewness and excess-kurtosis values of the predictor x. In other words, when Model 1 approximates the data-generating mechanism, y will be closer to the normal distribution than x. This asymmetry property opens the door for evaluation of the directional plausibility of a linear model by evaluating the skewness and excess kurtosis of a tentative response and a tentative predictor. Note that Eqs. 4 and 5, as proposed by Dodge and Rousson (2000, 2001), hold for the bivariate case. However, a two-step regression approach can be used to adjust for covariates defined above. First, two regression models are estimated in which x and y serve as responses and covariates z_j are used as independent variables—that is, $y = \sum_{j=1}^{J} b_{yz_j} z_j + e_{yz_j}$ and $x = \sum_{j=1}^{J} b_{xz_j} z_j + e_{xz_j}$. Next, the estimated regression residuals of the two models, e_{yz_j} and e_{xz_j}, are used as auxiliary variables reflecting the (unexplained) portion of variation after adjusting for the covariates z_j. Regressing e_{yz_j} on e_{xz_j} gives the same regression coefficient as obtained in the multiple linear model $\{x, z_j\} \rightarrow y$ and the OLS model $e_{yz_j} \rightarrow e_{xz_j}$ gives the identical estimate as in the multiple linear model $\{y, z_j\} \rightarrow x$. Direction dependence decisions can then be based on these auxiliary measures. For example, for one covariate z, Model 1 extends to

$$y = b_{yx}x + b_{yz}z + e_{y(xz)} \tag{6}$$

"Purified" measures of x and y are obtained through $e_{xz} = x - b'_{xz}z$ and $e_{yz} = y - b'_{yz}z$, where b'_{xz} and b'_{yz} denote the OLS estimates when regressing x and y on z. Then Eq. 6 can be rewritten as

$$e_{yz} = a_{yx}e_{xz} + \theta_{yx} \tag{7}$$

with $a_{yx} = b_{yx} = \left(\rho_{xy} - \rho_{yz}\rho_{xz}\right) / \left(1 - \rho_{xz}^2\right)$ being the partial regression coefficient and θ_{yx} denoting the error term that is identical to $e_{y(xz)}$. For the model in Eq. 7, one obtains (a proof is given in online Appendix A)

$$\rho_{xy|z}^3 = \frac{\gamma_{e_{yz}}}{\gamma_{e_{xz}}} \tag{8}$$

$$\rho_{xy|z}^4 = \frac{\kappa_{e_{yz}}}{\kappa_{e_{xz}}} \tag{9}$$

(as long as $\gamma_{e_{xz}} \neq 0$ and $\kappa_{e_{xz}} \neq 0$), where $\rho_{xy|z} = a_{yx}\frac{\sigma_{e_{xz}}}{\sigma_{e_{yz}}} = \frac{\rho_{xy} - \rho_{yz}\rho_{xz}}{\sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}}$ is the partial regression coefficient of x and y adjusting for z, and $\gamma_{e_{yz}}$, $\gamma_{e_{xz}}$, $\kappa_{e_{yz}}$, and $\kappa_{e_{xz}}$ are the skewness and excess-kurtosis values of e_{yz} and e_{xz}. Because $-1 \leq \rho_{xy|z} \leq 1$, higher moments of e_{yz} and e_{xz} possess the same properties as higher moments of x and y in the bivariate case. Under the model $x \rightarrow y$, we obtain $|\gamma_{e_{yz}}| < |\gamma_{e_{xz}}|$ and $|\kappa_{e_{yz}}| < |\kappa_{e_{xz}}|$, whereas $|\gamma_{e_{yz}}| > |\gamma_{e_{xz}}|$ and $|\kappa_{e_{yz}}| > |\kappa_{e_{xz}}|$ hold under model $y \rightarrow x$.

Presence of confounders Because any continuous nonnormal confounder can affect both the distribution of x and the distribution of y, directional decisions based on higher moments of x and y are influenced by (1) the magnitude of nonnormality of u, (2) the connection strength of u and x, and (3) the connection strength of u and y. Formally, this can be shown by rewriting higher moments of x and y as a function of higher moments of u—that is, $\gamma_x = \rho_{xu}^3\gamma_u$, $\kappa_x = \rho_{xu}^4\kappa_u$, $\gamma_y = \rho_{yu}^3\gamma_u$, and $\kappa_y = \rho_{yu}^4\kappa_u$ (which follows from applying Dodge and Rousson's, 2000, 2001, results to Model 3). Thus, the statistical power to determine the direction of effect depends on the magnitude of confounding effects and the degree of nonnormality of the confounder. If either the distribution of the confounder is close to normality or the influence of the confounder is weak (i.e., ρ_{xu} and ρ_{yu} are close to zero), no decisions can be made due to lack of sufficient nonnormality of x and y. The influence of the confounder on direction dependence decisions is given through

$$\frac{\gamma_y}{\gamma_x} = \left(\frac{\rho_{yu}}{\rho_{xu}}\right)^3 \tag{10}$$

and

$$\frac{\kappa_y}{\kappa_x} = \left(\frac{\rho_{yu}}{\rho_{xu}}\right)^4 \tag{11}$$

Thus, directional conclusions depend on the relative strength of the confounding effects. No biases in terms of model selection are expected when $|\rho_{yu}| < |\rho_{xu}|$ because $|\gamma_y| < |\gamma_x|$ and $|\kappa_y| < |\kappa_x|$ still hold, which suggests the model $x \rightarrow y$. In contrast, biases are likely to occur when $|\rho_{yu}| > |\rho_{xu}|$, because $|\gamma_y| > |\gamma_x|$ and $|\kappa_y| > |\kappa_x|$ increase the risk of erroneously selecting the mis-specified model $y \rightarrow x$.

Statistical inference von Eye and DeShon (2012) proposed using normality tests, such as D'Agostino's (1971) skewness

[3] Note that correlation and higher-moment parameters refer to population values, which implies that DDA quantities will exactly hold in the population. Due to sampling variability, DDA quantities will not hold exactly for sample estimates but converge to their true values with increasing sample size.

and/or Anscombe and Glynn's (1983) kurtosis test, to evaluate hypotheses compatible with observed-variable based direction dependence. Directional decisions are based on separately evaluating nonnormality of predictor and response. In addition, Pornprasertmanit and Little (2012) suggested nonparametric bootstrap CIs for higher-order moment differences ($\Delta(\gamma) = |\gamma_x| - |\gamma_y|$ and $\Delta(\kappa) = |\kappa_x| - |\kappa_y|$).

DDA Component II: Distributional properties of error terms

Absence of confounders The second DDA component focuses on the distributional shape of the error terms, e_{yx} and e_{xy}. In essence, distributional differences of the two error terms are likely to occur when the nonnormal "true" predictor is erroneously used as the outcome because predictor nonnormality will, to some extent, be preserved in the error term of the mis-specified model. Wiedermann, Hagmann, Kossmeier, and von Eye (2013), Wiedermann, Hagmann, and von Eye (2015), and Wiedermann (2015) showed that higher moments of the error term obtained from the mis-specified model (e_{xy}) can be expressed as functions of the third and fourth moments of the true predictor (x)—that is,

$$\gamma_{e_{xy}} = \left(1-\rho_{xy}^2\right)^{3/2}\gamma_x \tag{12}$$

and

$$\kappa_{e_{xy}} = \left(1-\rho_{xy}^2\right)^{2}\kappa_x \tag{13}$$

Thus, the skewness and excess kurtosis of e_{xy} systematically increase with the magnitude of nonnormality of the "true" predictor. Furthermore, because normality of the error term is assumed in the "true" model (i.e., $\gamma_{e_{yx}} = \kappa_{e_{yx}} = 0$), differences in higher moments of e_{yx} and e_{xy} provide, again, information about the directional plausibility of a linear model. This DDA component can straightforwardly be extended to multiple linear regression models when adjusting for possible covariates (cf. Wiedermann & von Eye, 2015b). Under model $x \rightarrow y$, one obtains $|\gamma_{e_{xy}}| > |\gamma_{e_{yx}}|$ and $|\kappa_{e_{xy}}| > |\kappa_{e_{yx}}|$; under model $y \rightarrow x$, one obtains $|\gamma_{e_{xy}}| < |\gamma_{e_{yx}}|$ and/or $|\kappa_{e_{xy}}| < |\kappa_{e_{yx}}|$.

Presence of confounders When an unmeasured confounder is present, the two competing models can be written as

$$y = b'_{yx}x + e'_{yx} \tag{14}$$

$$x = b'_{xy}y + e'_{xy} \tag{15}$$

where b'_{yx} and b'_{xy} are biased estimates of b_{yx} and b_{xy}. Although the causal estimate in Eq. 14 is biased, the model still correctly represents the data-generating process in terms of directionality. In this case, higher moments of e'_{yx} and e'_{xy} depend on the magnitude of nonnormality of u and the magnitudes of b_{xu} and b_{yu}. Specifically, the higher moments can be written as functions of semipartial correlations and higher moments of u. That is, for e'_{yx} one obtains

$$\begin{aligned}\gamma_{e'_{yx}} &= \rho^3_{y(u|x)}\gamma_u \\ \kappa_{e'_{yx}} &= \rho^4_{y(u|x)}\kappa_u\end{aligned} \tag{16}$$

and for e'_{xy} one obtains

$$\begin{aligned}\gamma_{e'_{xy}} &= \rho^3_{x(u|y)}\gamma_u \\ \kappa_{e'_{xy}} &= \rho^4_{x(u|y)}\kappa_u,\end{aligned} \tag{17}$$

with $\rho_{y(u|x)} = \left(\rho_{yu}-\rho_{xy}\rho_{xu}\right)/\sqrt{1-\rho^2_{xy}}$ being the semipartial correlation coefficient for y and u given x and $\rho_{x(u|y)} = \left(\rho_{xu}-\rho_{xy}\rho_{yu}\right)/\sqrt{1-\rho^2_{xy}}$ describing the semipartial correlation between x and u given y (see online Appendix A for a proof). The distribution of both error terms will be close to normality and, thus, no distinct decision is possible when u is close to normality and/or semi-partial correlations are close to zero. If the confounder is sufficiently nonnormal, the distributional properties of error terms and, thus, of directional decisions depend on the magnitude of the semipartial correlations. Unbiased directional decisions are possible when $|\rho_{y(u|x)}| < |\rho_{x(u|y)}|$ because $|\gamma_{e'_{xy}}| > |\gamma_{e'_{yx}}|$ and $|\kappa_{e'_{xy}}| > |\kappa_{e'_{yx}}|$, which implies $x \rightarrow y$. In contrast, if $|\rho_{y(u|x)}| > |\rho_{x(u|y)}|$, then erroneously selecting $y \rightarrow x$ is likely to occur because $|\gamma_{e'_{xy}}| < |\gamma_{e'_{yx}}|$ and $|\kappa_{e'_{xy}}| < |\kappa_{e'_{yx}}|$.

Statistical inference Again, nonnormality tests can be used to separately evaluate distributional properties of model residuals (cf. Wiedermann et al., 2015). An asymptotic significance test and bootstrap CIs for the skewness difference of residuals ($\Delta(\gamma_e) = |\gamma_{e_{xy}}| - |\gamma_{e_{yx}}|$) have been proposed by Wiedermann et al. (2015) and Wiedermann and von Eye (2015b). The asymptotic test requires normality of the "true" error term. Only error symmetry is required for the bootstrap approach. Analogous procedures for the difference in excess-kurtosis values were discussed by Wiedermann (2015).

DDA Component III: Independence properties of predictor and error term

Absence of confounders The independence assumption in the linear model implies that the magnitude of the error made when fitting the response is not related in any form to the predictor(s). In OLS regression, it is well-known that estimated residuals will be linearly uncorrelated with the predictor(s), which holds even when the model is directionally mis-specified. However, when the "true" predictor x is nonnormal, the error term and the predictor of the mis-specified model, y and e_{xy}, will be *stochastically nonindependent*. To illustrate this, we start with a simulated data example. Two variables (x and y) were generated according to the linear model $x \rightarrow y$ (with

zero intercept, unit slope, and a standard normal error term e_{yx}). The "true" predictor x was either drawn form a standard normal ($\gamma_x = \kappa_x = 0$), a standard uniform (i.e., $\gamma_x = 0$, $\kappa_x = -1.2$), or a chi-square distribution with eight degrees of freedom ($\gamma_x = 1$, $\kappa_x = 1.5$). Figure 2 shows scatterplots of the observed predictors and estimated residuals for the "true" model and the mis-specified model $y \rightarrow x$. In the normal case, the two models cannot be distinguished from each other. That is, for both models circular data patterns occur that can be expected due to linear uncorrelatedness. This no longer holds for nonnormal predictors. Here, the two competing models *are mutually distinguishable*. Although linear uncorrelatedness also holds for all nonnormal data scenarios, clear dependence structures occur in the mis-specified model. Note that these dependence structures are not the result of special properties of the uniform and the chi-square distributions. In fact, the opposite is the case. The normal distribution constitutes the special case in which competing models cannot be uniquely distinguished, because uncorrelatedness implies stochastic independence in the normal domain (cf. Hoyer, Shimizu, Kerminen, & Palviainen, 2008).

Formally, nonindependence in the mis-specified model becomes intuitively obvious if we solve for the error term of the mis-specified model in Eq. 2 and insert the "true" Model 1, which results in (see also Entner, Hoyer, & Spirtes, 2012; Shimizu, 2016)

$$e_{xy} = x - b_{xy}y = \left(1-\rho_{xy}^2\right)x - b_{xy}e_{yx} \tag{18}$$

Thus, both the "true" predictor x and the "true" error term e_{yx} contribute to y in Eq. 1 and e_{xy} in Eq. 18. Although this illustration serves as an intuitive explanation, a rigorous proof of nonindependence follows from the Darmois–Skitovich theorem (Darmois, 1953; Skitovich, 1953). The theorem states that if two linear functions (υ_1 and υ_2) of the same independent random variables w_j ($j = 1, \ldots, J$), $\upsilon_1 = \sum_j \alpha_j w_j$ and $\upsilon_2 = \sum_j \beta_j w_j$, with α_j and β_j being constants, are independent, then all w_j for which $\alpha_j\beta_j \neq 0$ must be normally distributed. The reverse corollary implies that if a common w_j exists that is nonnormal, then υ_1 and υ_2 must be nonindependent (cf. Shimizu et al., 2011; Wiedermann & von Eye, 2015a). Thus, e_{xy} in Eq. 18 and y in Eq. 1 are nonindependent because of the common nonnormal variable x, and $\left(1-\rho_{xy}^2\right)b_{yx} \neq 0$ (excluding $|\rho_{xy}| = 1$, due to practical irrelevance). Since the Darmois–Skitovich theorem applies for J variables, covariates can straightforwardly be included in the Models 1 and 2, provided that the covariates fulfill the requirements described above. However, the Darmois–Skitovich theorem concerns *continuous* random variables w_j. Thus, when categorical covariates exist, a two-step regression approach should be applied first with subsequent DDA being performed on residualized x and y variables. Because independence is assumed in the correctly specified model, direction dependence statements are possible through separately evaluating independence in competing models (cf. Shimizu et al., 2011; Wiedermann & von Eye, 2015a). In essence, if the null hypothesis $H_0 : x \perp e_{yx}$ is retained and, at the same time, $H_0 : y \perp e_{xy}$ is rejected, then it is more likely that the observed effect transmits from x to y. Conversely, if $H_0 : x \perp e_{yx}$ is rejected and $H_0 : y \perp e_{xy}$ is retained, then the model $y \rightarrow x$ should be preferred.

Presence of confounders When confounding affects the relation between x and y, predictor(s) and errors of *both* models

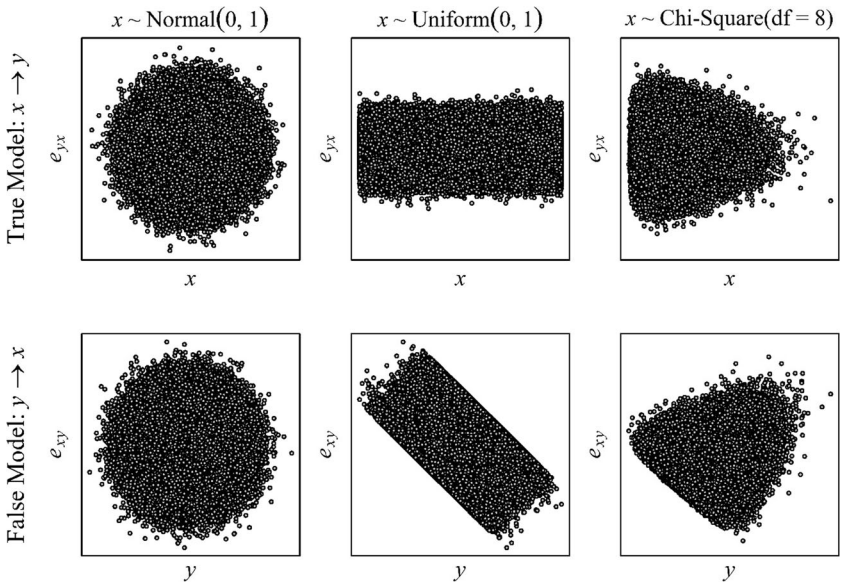

Fig. 2 Bivariate patterns of the predictors and residuals of competing linear models ($x \rightarrow y$ and $y \rightarrow x$) for both normal and nonnormal "true" predictors

contain information of the confounder. This can be shown by solving Eqs. 14 and 15 for the corresponding error terms, which gives

$$e'_{yx} = \left[b_{yu} + \left(b_{yx} - b'_{yx}\right)b_{xu}\right]u + \left(b_{yx} - b'_{yx}\right)e_{xu} + e_{yu} \quad (19)$$

$$e'_{xy} = \left[b_{xu} - b'_{xy}\left(b_{yu} + b_{yx}b_{xu}\right)\right]u + \left(1 - b'_{xy}b_{yx}\right)e_{xu} + b'_{xy}e_{yu} \quad (20)$$

Thus, through reconsidering the "true" model given in Eq. 3 and, again, making use of the Darmois–Skitovich theorem, one concludes that the independence assumption is likely to be violated in *both* candidate models whenever a nonnormal confounder is present and $\left[b_{yu} + \left(b_{yx} - b'_{yx}\right)b_{xu}\right]b_{xu}$ and $\left[b_{xu} - b'_{xy}\left(b_{yu} + b_{yx}b_{xu}\right)\right]b_{yu}$ deviate from zero.

Statistical inference Significance tests to evaluate nonindependence of (linearly uncorrelated) variables have extensively been discussed in signal processing (Hyvärinen, Karhunen, & Oja, 2001). The first class of tests considered here uses the basic definition of stochastic independence, $E[g_1(\upsilon_1)g_2(\upsilon_2)] - E[g_1(\upsilon_1)]E[g_2(\upsilon_2)] = 0$ for any absolutely integrable functions g_1 and g_2. Thus, independence tests can be constructed using correlation tests of the form cor[$g_1(x)$, $g_2(e_{yx})$] and cor[$g_1(y)$, $g_2(e_{xy})$] , where at least one function is nonlinear. These tests are easy to use because they essentially rely on the Pearson correlation test applied to nonlinearly transformed variables.

Two nonlinear functions may be of particular value in the present context: the square function, $g(\upsilon) = \upsilon^2$, and the hyperbolic tangent function, $g(\upsilon) = \tanh(\upsilon)$. The square function constitutes a powerful candidate because covariances of predictor and error in the mis-specified model contain information of higher moments of the "true" predictor (Wiedermann and von Eye, 2015a)—that is,

$$\operatorname{cov}\left(y, e_{xy}^2\right) = \rho_{xy}\left(1 - \rho_{xy}^2\right)^2\gamma_x \quad (21)$$

$$\operatorname{cov}\left(y^2, e_{xy}\right) = \rho_{xy}^2\left(1 - \rho_{xy}^2\right)\gamma_x \quad (22)$$

$$\operatorname{cov}\left(y^2, e_{xy}^2\right) = \rho_{xy}^2\left(1 - \rho_{xy}^2\right)^2\kappa_x \quad (23)$$

In other words, the power of detecting nonindependence in $y \rightarrow x$ increases with nonnormality of x. Although proofs for Eqs. 21 and 22 can be found in Wiedermann and von Eye (2015a, 2016), a proof of Eq. 23 is given in online Appendix A. Note that the covariances in Eqs. 21 and 23 involve squared residuals, which reveals a direct link to significance tests originally designed for detecting patterns of heteroscedasticity (cf. Wiedermann, Artner, & von Eye, 2017). Because heteroscedasticity occurs, among others, when the variance of the error can be expressed as a function, g, of independent variables—that is, $Var\left(e_{xy}|y\right) = \sigma^2_{e_{xy}}g(y)$ (see, e.g., Kaufman, 2013), it follows that homoscedasticity tests that relate squared residuals to functions of model predictors (such as the Breusch–Pagan test), are likely to indicate patterns of nonconstant error variances in directionally mis-specified models.

Covariances based on the hyperbolic tangent function have been proposed by Hyvärinen (2010) and Hyvärinen and Smith (2013). The value $\tanh(\upsilon)$ is the derivative of the log-density of an inverse hyperbolic cosine distribution that provides an approximation of the likelihood ratio of directionally competing models in the bivariate case. The inverse hyperbolic cosine distribution constitutes a reasonably close approximation for several leptokurtic observed variables (cf. Mumford & Ramsey, 2014). Although tanh-based correlation tests are ideally suited for symmetric nonnormal variables, the statistical power of this approach can be expected to be low for skewed variables.

Since the choice of g_1 and g_2 is almost arbitrary, nonlinear approaches do not constitute rigorous directionality tests. Additional Type II errors beyond cases of small sample sizes are introduced because testing all existing functions g_1 and g_2 is impossible. Recently, a promising alternative was suggested, the Hilbert-Schmidt Independence Criterion (HSIC; Gretton et al., 2008). The HSIC evaluates the independence of functions of random variables and is provably omnibus in detecting any dependence between two random variables in the large sample limit. Sen and Sen (2014) introduced the HSIC in the context of testing the independence of predictors and error terms of linear regression models and proposed a bootstrap approach to approximating the distribution of the test statistic.

Model selection

Reconsidering the possible outcomes for the three DDA components, it becomes evident that each model in Fig. 1 can be uniquely identified through specific DDA-component patterns. Table 1 summarizes these patterns, together with corresponding significance tests for each DDA component. In general, DDA model selection requires the specification of a *target* and an *alternative* model. The selection of whether $x \rightarrow y$ or $y \rightarrow x$ serves as the target model is arbitrary in terms of model comparison. However, following the logic of confirmatory model testing, we suggest that the target model reflect the substantive causal theory of interest and that the alternative model reflect the contradicting theory. The target model, for example $x \rightarrow y$, finds support when (1) the distribution of the response y is closer to normality than the distribution of x, (2) the residual distribution of $x \rightarrow y$ is closer to normality than the residuals of $y \rightarrow x$, and (3) the independence assumption of

Table 1 Properties, significance tests, and patterns of DDA components for the three candidate models

	Distribution of Observed Variables	Distribution of Error Terms	Independence of Predictor and Error Term
General Properties	The "true" outcome of a confounder-free model will always be closer to the normal distribution than the true predictor	The "true" error term of a confounder-free model will always be closer to the normal distribution than the error term of the mis-specified model	Predictor and error term of the "true" model will be independent, whereas nonindependence will be observed in the mis-specified model
Significance Tests	Separate D'Agostino skewness and Anscombe–Glynn kurtosis tests	Separate D'Agostino skewness and Anscombe–Glynn kurtosis tests	Separate nonlinear correlation tests Separate Breusch–Pagan and
	Bootstrap confidence interval for higher-moment differences	Asymptotic skewness and excess-kurtosis difference tests Bootstrap confidence interval for higher-moment differences	robust Breusch–Pagan homoscedasticity tests HSIC test
Model: $x \rightarrow y$	y is closer to the normal distribution than x: $\lvert\gamma_y\rvert < \lvert\gamma_x\rvert$ $\lvert\kappa_y\rvert < \lvert\kappa_x\rvert$	e_{yx} is closer to the normal distribution than is e_{xy}: $\lvert\gamma_{e_{yx}}\rvert < \lvert\gamma_{e_{xy}}\rvert$ $\lvert\kappa_{e_{yx}}\rvert < \lvert\kappa_{e_{xy}}\rvert$	e_{yx} and x are independent, and e_{xy} and y are dependent
Model: $y \rightarrow x$	x is closer to the normal distribution than y: $\lvert\gamma_x\rvert < \lvert\gamma_y\rvert$ $\lvert\kappa_x\rvert < \lvert\kappa_y\rvert$	e_{xy} is closer to the normal distribution than e_{yx}: $\lvert\gamma_{e_{xy}}\rvert < \lvert\gamma_{e_{yx}}\rvert$ $\lvert\kappa_{e_{xy}}\rvert < \lvert\kappa_{e_{yx}}\rvert$	e_{xy} and y are independent, and e_{yx} and x are dependent
Presence of Confounder:	Higher moment differences depend on the correlations between x and u and y and u: • $\lvert\gamma_y\rvert < \lvert\gamma_x\rvert$ and $\lvert\kappa_y\rvert < \lvert\kappa_x\rvert$ if $\lvert\rho_{xu}\rvert > \lvert\rho_{yu}\rvert$ • $\lvert\gamma_x\rvert < \lvert\gamma_y\rvert$ and $\lvert\kappa_x\rvert < \lvert\kappa_y\rvert$ if $\lvert\rho_{xu}\rvert < \lvert\rho_{yu}\rvert$	Higher-moment differences of e_{yx} and e_{xy} depend on the semipartial correlation coefficients $\rho_{y(u\mid x)}$ and $\rho_{x(u\mid y)}$: • $\lvert\gamma_{e_{yx}}\rvert < \lvert\gamma_{e_{xy}}\rvert$ and $\lvert\kappa_{e_{yx}}\rvert < \lvert\kappa_{e_{xy}}\rvert$ if $\lvert\rho_{y(u\mid x)}\rvert < \lvert\rho_{x(u\mid y)}\rvert$ • $\lvert\gamma_{e_{xy}}\rvert < \lvert\gamma_{e_{yx}}\rvert$ and $\lvert\kappa_{e_{xy}}\rvert < \lvert\kappa_{e_{yx}}\rvert$ if $\lvert\rho_{x(u\mid y)}\rvert < \lvert\rho_{y(u\mid x)}\rvert$	Independence assumption will be violated in both models—that is, e_{yx} and x are dependent, and e_{xy} and y are also dependent

residuals and predictor(s) holds for $x \rightarrow y$ and is, at the same time, violated for model $y \rightarrow x$. Here, DDA's independence procedures constitute the key element to test for the presence of unobserved confounders. In this case, independence *must hold* for $x \rightarrow y$ and the independence assumption *must be violated* for $y \rightarrow x$ in order to conclude that an effect is transmitted from x to y. Otherwise, one has to conclude that unmeasured confounders are present whenever the independence assumption is either violated or satisfied in both models (the latter possibility emerges from the fact that confounders can decrease the skewness/excess kurtosis of x and y to a degree that renders nonindependence no longer detectable). When independence tests allow a distinct decision, higher-moment tests for the observed variables and residuals are then used to test the directionality assumption inherent to the target model.

A worked empirical example with SPSS

To make DDA accessible to applied researchers, we provide three SPSS macros. `DDA_VarDist` analyzes the distributional properties of the variables, `DDA_ResidDist` evaluates the distributional properties of the residuals, and `DDA_Indep` implements tests to evaluate the independence assumption regarding the competing candidate models (the DDA macros and introductory material can be downloaded from [http://www.ddaproject.com/]). All macros make use of positional arguments to control the input parameters. Table 2 summarizes the input parameters and options for the three macros, together with generic examples of their use. Computational details and introductions into implemented DDA inference methods are given in online Appendix B.

To present a fully worked empirical example (consisting of preevaluating distributional requirements, building a valid target model, and subsequently using DDA) and demonstrate the use of the macros for DDA, we use data from a cross-sectional study on the triple-code model (Dehaene & Cohen, 1998). The triple-code model is used to explain the development of numerical cognition in children and proposes that numbers are represented in three different codes that serve different purposes in number processing. The *analog magnitude code* (AMC) represents numbers on a mental number line, includes knowledge of the proximity and size of numerical quantities, and is used in approximate estimations and magnitude comparisons. The *auditory verbal code* (AVC) represents numbers in syntactically organized word sequences that are important for verbal input/output, counting, and retrieving memorized arithmetic facts. The *visual Arabic code* (VAC) represents numerical quantities in Arabic format necessary for multidigit operations and parity judgments. Using the triple code model, von Aster and Shalev (2007) suggested a hierarchical developmental model of numerical cognition in which AMC is

Table 2 Summary of arguments and their position in the three SPSS macros

Position	DDA_VarDist	DDA_ResidDist	DDA_Indep
1	Response variable	Response variable	Response variable
2	Predictor variable	Predictor variable	Predictor variable
3	Number of bootstrap samples for constructing nonparametric bootstrap CIs of higher moment differences	Number of bootstrap samples for constructing nonparametric bootstrap CIs of higher moment differences	Test of independence. Takes either integer values representing the power (e.g. `2` represents the square function) or the characters "`T`" (for tanh), or "`H`" (HSIC test).
4	Confidence level in relative frequencies	Confidence level in relative frequencies	(*) `missing=listwise` or "`pairwise`", indicating whether listwise or pairwise deletion is used for missing values (default: `missing=listwise`).
5	Direction of the alternative hypotheses. Takes argument "`two.sided`", "`greater`" or "`less`".	Direction of the alternative hypotheses. Takes argument "`two.sided`", "`greater`" or "`less`".	(*) `save=` An indicator taking values "`T`" (true) or "`F`" (false) controlling whether estimated residuals should be saved in the data file (default: `save="F"`).
6	(*) `missing=listwise` or "`pairwise`", indicating whether listwise or pairwise deletion is used for missing values (default: `missing=listwise`).	(*) `missing=listwise` or "`pairwise`", indicating whether listwise or pairwise deletion is used for missing values (default: `missing=listwise`).	(*) `B=` number of bootstrap samples for HSIC test.
7	(*) `save=` An indicator taking values "`T`" (true) or "`F`" (false) controlling whether estimated residuals should be saved in the data file (default: `save="F"`).	(*) `save=` An indicator taking values "`T`" (true) or "`F`" (false) controlling whether estimated residuals should be saved in the data file (default: `save="F"`).	(*) `cov=` Optional list of covariate(s). The end of command indicates the end of list.
8	(*) `cov=` Optional list of covariate(s). The end of command indicates the end of list.	(*) `cov=` Optional list of covariate(s). The end of command indicates the end of list.	–

Example:	`DDA_VarDist y x 500 .95`
	`two.sided`
	`missing=pairwise`
	`save=T cov=z`
		`DDA_ResidDist y x 500 .95`
		`two.sided`
		`cov=z`
			`DDA_Indep y x H`
			`save=T`
			`B=500`
			`cov=z`

In the generic example, `y` is the outcome, `x` is the predictor, and `z` is a covariate. Arguments marked with (*) are optional

viewed as an inherited core system necessary to further develop the AVC and, as well, the VAC. In other words, the model posits a directional link between AMC and AVC (i.e., AMC → AVC) and AMC and VAC (i.e., AMC → VAC). In the present demonstration, we focus on the directionality of AMC and AVC.

Koller and Alexandrowicz (2010) collected AMC and AVC ability measures for 341 second- to fourth-grade elementary school children (185 girls and 156 boys, aged between 6 and 11 yrs.) using the Neuropsychological Test Battery for Number Processing and Calculation in Children (ZAREKI-R; von Aster, Weinhold Zulauf, & Horn, 2006). AMC sum scores are based on 31 dichotomous items (focusing on perceptual quantity estimation, placing numbers on an analog number line, counting backward, enumeration, magnitude comparison of spoken numbers,

and contextual magnitude judgment), and AVC sum scores are based on 52 dichotomous items (mental calculations [addition, subtraction, and multiplication], repeating numbers forward and backward, and story problems). The sum scores were standardized prior to the analysis in order to improve interpretability. Because fourth-graders were most likely to solve all items of the AMC scale, we focused on the second- and third-grade children (n = 216; 123 girls and 93 boys) in order to avoid biased DDA results due to ceiling effects.

DDA was used to evaluate the two competing regression models (AMC → AVC vs. AVC → AMC) under adjustment for the covariates age (variable `age`), time needed for test completion in minutes (as an indirect measure of the perceived test difficulty; variable `time`), and preexisting difficulties with numerical quantities (0 = no, 1 = yes; variable `diff`). All covariates preceded test performance in time, and we could exclude cyclic relations. Table 3 shows pairwise correlations and descriptive measures for all considered variables. Before applying DDA, two preevaluation stages are crucial to obtaining meaningful results: (1) evaluation of the distributional requirements for DDA and (2) carefully building a valid target model. Both stages are discussed in detail below.

Distributional requirements for DDA

DDA requires that the distributions of the observed variables deviate from normality. Thus, before estimating the target model (AMC → AVC), we evaluated the assumption of nonnormality of the variables. The AMC and AVC measures were negatively skewed, with excess-kurtosis values greater than zero (Table 3). The Shapiro–Wilk test rejected the null hypothesis of normality for both ability measures (ps < .001). Visual inspection was used to rule out the presence of outliers, and frequencies of the minimum/maximum scores were computed in order to assess potential floor/ceiling effects. For the AVC scale, no participant reached the minimum or maximum score. For AMC, no participants received the minimum, and 14 out of 216 (6.5%) reached the maximum score, which is clearly below the commonly used cutoff of 15%–20% to define a ceiling effect (e.g., Lim et al., 2015; Terwee et al., 2007). Overall, the variables can be considered in line with the distributional requirements of DDA.

Estimating and validating the target model

We started by partialling out the effect of the binary indicator "difficulties with numbers" using separate OLS regressions and extracted residuals as "purified" AMC and AVC measures—that is, $\texttt{AMC\_r} = \texttt{AMC} - (0.032 - 0.850 \times \texttt{diff})$ and $\texttt{AVC\_r} = \texttt{AVC} - (0.035 - 0.795 \times \texttt{diff})$. Figure 3 shows the univariate distributions and the bivariate scatterplot (with the LOWESS smoothed line superimposed) for `AMC_r` and `AVC_r`. In a similar fashion, we partialled out the effect of the binary indicator on the remaining continuous covariates —that is, $\texttt{age\_r} = \texttt{age} - (8.222 + 0.111 \times \texttt{diff})$, and $\texttt{time\_r} = \texttt{time} - (28.411 + 2.535 \times \texttt{diff})$.

Next, we estimated the target model (`AMC_r` → `AVC_r`) under adjustment for the continuous covariates (`time_r` and `age_r`) and evaluated the validity of the model using regression diagnostics. Table 4 summarizes the results for both the target model (upper panel) and the alternative model (lower panel). The linearity assumption of the target model was confirmed through inspection of the LOWESS plots and inclusion of higher-order terms. Adding quadratic terms for the predictors did not significantly improve the model fit (e.g., including squared values of `AMC_r` increased the model R^2 from .507 to .510, which was nonsignificant on the basis of a 5% significance level). Variance inflation factors for the predictors varied from 1.16 to 1.27, suggesting the absence of multicollinearity issues. Furthermore, we estimated leverage values and Cook's distances for the model, to check for the presence of highly influential observations. We excluded one observation from the subsequent DDA with a maximum Cook's distance of .195 (95% of the observations had a Cook's distance smaller than or equal to .023) and a leverage value of 0.075, which exceeded three times the average leverage values.

Table 3 Bivariate Pearson correlations and descriptive measures of observed variables (means and standard deviations of AMC and AVC are based on sum scores)

	Variable	(2)	(3)	(4)	(5)	*M*	*SD*	γ	κ
(1)	Analogue magnitude code (AMC)	.725	.240	– .442	– .374	25.47	4.26	– 1.13	1.18
(2)	Auditory verbal code (AVC)	–	.294	– .466	– .364	36.08	6.75	– 0.72	0.45
(3)	Years of age		–	– .304	.067	8.26	0.79	0.13	– 0.35
(4)	Time to complete the test (in minutes)			–	.198	29.29	6.12	0.80	0.60
(5)	Preexisting difficulties				–	0.35	0.48	0.64	– 1.59

Evaluating the direction of effect

To test whether the target model was indeed better-suited to approximate the data-generating mechanism, we first applied the DDA_VarDist macro. After setting the working directory C:\myproject (used to save and read temporary files during computations) and reading and activating the dataset data_example.sav using the code

```
CD "C:\myproject".
GET FILE = "C:\myproject\data_example.sav".
DATASET NAME data WINDOW = FRONT.
DATASET ACTIVATE data.
```

the following command gives the results for observed-variable-based direction dependence tests using 1,000 bootstrap samples (used for the construction of confidence interval [CI] limits), a confidence level of 95%, and two-sided significance tests of normality:

```
DDA_VarDist AVC_r AMC_r 1000 0.95 two.sided cov = age_r time_r
```

The corresponding output is given in Box 1. The upper panel summarizes the results of D'Agostino skewness and Anscombe–Glynn kurtosis tests for the putative response (columns 1–3) and predictor (columns 4–6). Skewness and excess-kurtosis values were close to zero for AVC_r and we can retain the null hypothesis of normality. In contrast, AMC_r significantly deviated from normality with respect to skewness. The results for excess-kurtosis estimates point in the same direction. The lower panel of Box 1 reports the 95% nonparametric bootstrap CIs for the differences in skewness $\Delta(\gamma) = |\gamma_{AMC\_r}| - |\gamma_{AVC\_r}|$ and excess kurtosis $\Delta(\kappa) = |\kappa_{AMC\_r}| - |\kappa_{AVC\_r}|$. Although AMC_r is significantly more skewed than AVC_r, the difference in excess kurtosis was nonsignificant. Overall, the third-moment estimates provide evidence in line with direction dependence requirements necessary for AMC_r → AVC_r.

Box 1. Results from DDA_VarDist

```
Skewness and Excess Kurtosis Tests
        p_AVC_r  z-value      Sig  p_AMC_r  z-value      Sig
Skew     -.2958  -1.7935    .0729   -.8369  -4.5810    .0000
Exkurt    .3868   1.2638    .2063    .8389   2.1568    .0310

Bootstrap Confidence Intervals (Skewness and Kurtosis Difference)
       BootCILo BootCIUp
Skew      .2035    .8561
Exkurt   -.1165   1.8255
```

Note: The prefix "p_" is automatically added to variable names to indicate that covariates have been partialled out before performing higher moments tests.

Next, we evaluated the properties of the residuals obtained from the two competing models. The following command performs residual-based direction dependence tests (again, using two-sided tests, 1,000 bootstrap samples for constructing CI limits, and a 95% confidence level):

```
DDA_ResidDist AVC_r AMC_r 1000 0.95 two.sided cov = age_r time_r
```

The upper panel of Box 2 summarizes separate skewness and excess-kurtosis tests of the regression residuals. Columns 1–3 refer to the target model, and columns 4–6 give the results for the alternative model. Although the higher-moment estimates were larger (in absolute values) for the alternative model, we cannot the reject the null hypothesis of normality at the 5% level. Similar results were obtained for the higher-moment difference measures $\Delta(\gamma_e) = |\gamma_e^{(AVCr \to AMCr)}| - |\gamma_e^{(AMCr \to AVCr)}|$ and $\Delta(\kappa_e) = |\kappa_e^{(AVCr \to AMCr)}| - |\kappa_e^{(AMCr \to AVCr)}|$ (see the lower panel of Box 2). Both the asymptotic higher-moment difference tests (columns 1–3) and 95% nonparametric bootstrap CIs (last two columns) suggested that the two models are not distinguishable in terms of their residual distributions. Thus, no clear-cut decision is possible for this component.

Box 2. Results from DDA_ResidDist.

```
Skewness and Excess Kurtosis Tests
        r_AVC_r  z-value      Sig  r_AMC_r  z-value      Sig
Skew     .1037    .6391    .5228   -.3182  -1.9241    .0543
Exkurt  -.1061   -.1150    .9084   -.1747   -.3596    .7191

Skewness and Excess Kurtosis Difference Tests
          Diff.  z-value      Sig BootCILo BootCIUp
Skew      .2146   1.6533    .0983   -.0656    .5883
Exkurt    .0686    .1842    .8539   -.4020    .5349
```

Note: The prefix "r_" is automatically added to the variable names to indicate that higher moments tests are performed on extracted model residuals.

In the final step, we analyzed the independence properties of the two candidate models, which is the most important element for interpreting OLS estimates as causal. The command

```
DDA_Indep AVC_r AMC_r 2 cov = age_r time_r
```

computes Breusch–Pagan (BP) homoscedasticity tests and nonlinear correlation tests using the square function. The results are summarized in Box 3. The upper section of the output gives the results for the BP and the robust-BP tests for the Target Model and the Alternative Model. Overall, the results were clearly in favor of the target model—that is, the homoscedasticity assumption holds for the target model and, at the same time, is violated for the alternative model. In addition, we used scatterplots of the standardized predicted values and standardized residuals for both models as a visual aid (see Fig. 4). No conspicuous patterns were observed for the target model, whereas the plot for the alternative model suggested an inverse U-shaped pattern. The lower section of Box 3 summarizes the results of nonlinear correlation tests. In general, given a selected function g, Pearson correlation coefficients, t values, and p values are computed for cor[g(pred), e], cor[pred, g(e)], and cor[g(pred), g(e)]. In the present example, nonlinear correlation tests based on the square function again clearly favored the target model; that is, all tests were nonsignificant for the target model, and at the same time, all tests rejected the null hypothesis for the alternative model.[4] Finally, the command

Box 3. Results from DDA_Indep using homoscedasticity tests and nonlinear correlation tests based on the square function

```
Tests of Heteroscedasticity: Target Model
              Chisq         df        Sig
BP-test      1.5029     3.0000      .6816
robustBP     1.5871     3.0000      .6623

Tests of Heteroscedasticity: Alternative Model
              Chisq         df        Sig
BP-test     13.1779     3.0000      .0043
robustBP    14.4391     3.0000      .0024


Target Model: Predictor_Residual Non-linear Correlation Test
         est.corr    t-value         df        Sig
AM2_R      -.0682     -.9974   213.0000      .3197
AM_R2      -.0631     -.9226   213.0000      .3573
AM2_R2      .0281      .4096   213.0000      .6825

Alternative Model: Predictor_Residual Non-linear Correlation Test
         est.corr    t-value         df        Sig
AV2_R      -.1958    -2.9141   213.0000      .0039
AV_R2      -.2535    -3.8254   213.0000      .0002
AV2_R2      .1857     2.7585   213.0000      .0063
```

Note: The selected non-linear function will be added to truncated variable names [e.g., AV2_R = $cor(AVC^2, error)$, AV_R2 = $cor(AVC, error^2)$, and AV2_R2 = $cor(AVC^2, error^2)$].

```
DDA_Indep AVC_r AMC_r H B = 500 cov = age_r time_r
```

[4] We do not focus on tanh-based tests because of low power for asymmetrically distributed variables.

computes HSIC tests for the two competing models using 500 bootstrap samples. Box 4 gives the corresponding output. Note that the HSIC will be zero if and only if the predictor and the error term are stochastically independent. Again, a nonsignificant result was observed for the `Target Model`, whereas the HSIC reached significance for the `Alternative Model`. In sum, all independence measures indicated that `AMC_r` → `AVC_r` is more likely to hold for the present dataset.

Box 4. Results from `DDA_Indep` using the HSIC test

```
Target Model: Predictor_Residual Hilbert-Schmidt Independence Criterion
           HSIC        Sig
AM_R      .0452      .7020

Alternative Model: Predictor_Residual Hilbert-Schmidt Independence Criterion
           HSIC        Sig
AV_R      .5005      .0000
```

Considering the overall results of DDA for the numerical-cognition example, we conclude that, taking into account the covariates, AVC is indeed more likely to reflect the response, and AMC is more likely to be on the explanatory side. In other words, on the basis of the present sample, the DDA results empirically support von Aster and Shalev's (2007) hierarchical developmental model of numerical cognition.

Discussion

DDA allows researchers to test hypotheses compatible with the directional relation between pairs of variables while adjusting for covariates that possibly contribute to the causal process. This empirical falsification approach is based on the translation of a substantive causal theory into a linear target model that is then compared with the corresponding alternative model. The two models differ in the direction that is hypothesized for the causal process. DDA component patterns can then be used to either retain the target model, retain the directionally competing model, or conclude that no distinct decisions are possible due to the presence of unmeasured confounders. Here, it is important to reiterate that directional conclusions derived from DDA component patterns are based on the operationalization of latent constructs $\mathcal{X}$ and $\mathcal{Y}$ using the linear model as an approximation of an unknown "true" functional relation $\mathcal{F}$. Trustworthiness of DDA, thus, ultimately depends on both, the quality of operationalization and the validity of the linear model for the description of the causal mechanism. Although both requirements essentially apply to any linear modeling approach, they deserve particular attention in the context of DDA.

Because higher moments of variables constitute the key elements to select directionally competing models, DDA assumes that nonnormality of variables reflects inherent distributional characteristics of the constructs under study.

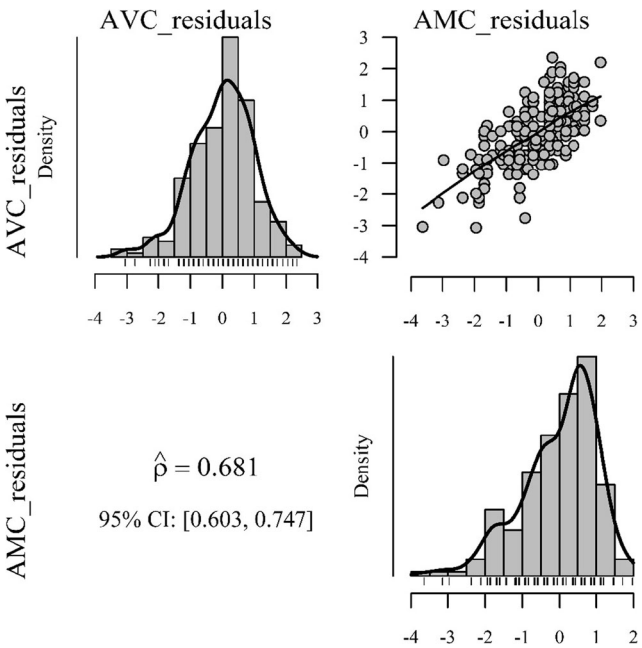

Fig. 3 Univariate distributions (main diagonal) and scatterplot (upper right panel) of "purified" analog magnitude code (AMC_residuals) and auditory verbal code (AVC_residuals) scores, including the bivariate LOWESS smoothed line for 216 second- and third-grade elementary school children

Table 4 Results of the two competing models (B = unstandardized coefficients, Std. Error = standard error, Beta = standardized coefficients)

	B	Std. Error	Beta	*t*-value	*p*-value
Response: AVC_r (adj. R^2 = 0.500)					
Intercept	<0.001	0.047	–	0.00	1.000
AMC (`AMC_r`)	0.563	0.052	0.583	10.86	<0.001
Years of age (`age_r`)	0.155	0.064	0.126	2.42	0.016
Time to complete test (`time_r`)	–0.025	0.009	–0.155	–2.85	0.005
Response: AMC_r (adj. R^2 = 0.473)					
Intercept	<0.001	0.050	–	0.00	1.000
AVC (`AVC_r`)	0.636	0.059	0.614	10.86	<0.001
Years of Age (`age_r`)	0.044	0.069	0.034	0.64	0.525
Time to complete test (`time_r`)	–0.022	0.009	–0.129	–2.30	0.022

Although the phenomenon of nonnormal variable distributions and its occurrence in practice have extensively been studied in the psychometric literature (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Cain, Zhang, & Yuan, 2017; D. L. Cook, 1959; Lord, 1955; Micceri, 1989), not every form of nonnormality makes variables eligible for DDA. In classical test theory, for example, the impact of discrimination and difficulty of a measurement instrument on the relation between latent traits and true score is well understood. To ensure that the observed score distributions adequately reflect distributional properties of a latent trait, the test characteristic curve should go straight through the range of the trait distribution, which is usually achieved by using items with a broad range of difficulties (Lord & Novick, 1968, p. 392). In addition, item response theory (IRT) models[5] such as the Rasch model (Rasch, 1960/1980, for dichotomous data) and the partial credit model (Masters, 1982, for polytomous data) are valuable alternatives. These models (1) come with empirical measures to evaluate the adequacy of describing a given dataset, (2) provide accordingly "weighted" parameter estimates (i.e., taking into account item difficulties), and (3) if the measurement model holds, exhibit the feature of *specific objectivity* (i.e., items can be compared irrespective of the distribution of person parameters and subjects can be compared using any proper set of items), which allows the most adequate estimation of the underlying trait distributions. For example, data on numerical cognition used for illustrative purposes were shown to be in line with the Rasch model (see Koller & Alexandrowicz, 2010), which implies that raw scores are sufficient statistics for the latent person abilities. In contrast, applying DDA in cases in which nonnormality of variables is a by-product of poor item selection, scaling (Ho & Yu, 2015), or the result of ceiling/floor effects will lead to biased results (note that, in the empirical example, the fourth grade children who were most likely to solve all scale-specific items were excluded to reduce the risk of biases due to ceiling effects). Overall, selecting high-quality measurement instruments at the study planning stage, or carefully evaluating

[5] We thank one of the anonymous reviewers for this suggestion.

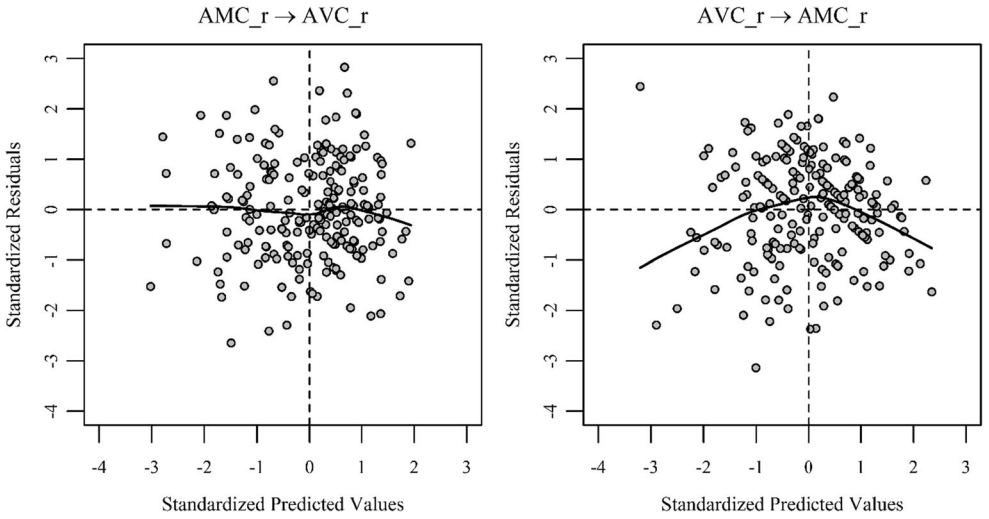

Fig. 4 Residual plot for the target (AMC_r → AVC_r) and the alternative (AVC_r → AMC_r) models, with bivariate LOWESS smoothed lines

psychometric properties of secondary data are central steps toward meaningful DDA outcomes.

Explanatory modeling, in general, requires that selected statistical models *f* can easily be linked to the corresponding theoretical model $\mathcal{F}$ (Shmueli, 2010). Because the "true" data-generating mechanism $\mathcal{Y} = \mathcal{F}(\mathcal{X})$ is unknown in any statistical modeling approach (Cudeck & Henly, 2003) empirical examinations whether $y = f(x)$ is close enough to $\mathcal{Y} = \mathcal{F}(\mathcal{X})$ are impossible. Appropriateness of *f* must be established indirectly through critically evaluating model validity using regression diagnostics (cf. Belsley, Kuh, & Welsch, 1980; R. D. Cook & Weisberg, 1982). Several model checks are indispensable before applying DDA. First, one needs to ensure that the assumption of linearity is justified (in the illustrative example we used visual diagnostics and evaluated changes in R^2 values when adding higher polynomials of all continuous variables). Second, evaluating potential issues of multicollinearity (e.g., inspecting pairwise predictor correlations and VIFs) are necessary to avoid biased inference due to inflated standard errors. Third, absence of outliers and highly influential data points must be confirmed (e.g., via examining Cook's distances, leverage statistics, or deleted studentized residuals). Ideally, the process of building a valid target model and the subsequent evaluation of its directional properties constitute two separate steps. This implies that unintended DDA outcomes should not be used as a basis to delete "misbehaving" data points.

The case of nonnormal "true" errors

The DDA framework presented here assumes that the "true" error follows a normal distribution. Although, in best practice applications, normality of residuals should routinely be evaluated to guarantee valid statistical inference (Box & Watson, 1962; Hampel, 1973; Pearson, 1931; White & MacDonald, 1980), normality is not required for OLS coefficients to be the best linear unbiased estimates. Normal "true" errors are particularly important for residual-distribution-based DDA tests when measures of both, skewness and excess kurtosis, are considered simultaneously because normality of the correctly specified error then serves as a benchmark for model comparison. However, when one only focuses on the skewness of competing error terms, model selection can be performed as long as $\gamma_{e_{yx}} = 0$—that is, no explicit assumptions are made concerning $\kappa_{e_{yx}}$. Model selection should then be based on nonparametric bootstrap CIs of skewness differences instead of the asymptotic skewness difference test (cf. Wiedermann & von Eye, 2015c). Reversely, when solely focusing on the excess kurtosis of error terms, no explicit assumptions are made concerning symmetry of the "true" error distribution and, as long as $\kappa_{e_{yx}} = 0$ holds for the "true" model, $\gamma_{e_{yx}}$ is allowed to vary within the range $-\sqrt{2}$ to $\sqrt{2}$ according to the skewness–kurtosis inequality $\kappa \geq \gamma^2 - 2$ (cf. Teuscher & Guiard, 1995).

Although DDA, based on the skewness and excess kurtosis of the observed variables, also requires normality of the "true" errors, focusing on either skewness or the excess kurtosis relaxes distributional assumptions about the "true" error in the same fashion (for a detailed discussion on distinguishing directionally competing models under error nonnormality, see also Wiedermann & Hagmann, 2015). In addition, alternative DDA measures based on higher-order correlations $\rho_{ij}(x,y) = \mathrm{cov}_{ij}(x,y)/\left(\sigma_x^i \sigma_y^j\right)$ with $\mathrm{cov}_{ij}(x,y) = E[(x - E[x])^i (y - E[y])^j]$ are available that do not make any assumptions about the "true" error distribution. Dodge and Rousson (2001)

showed that $\rho_{xy} = \rho_{12}(x, y)/\rho_{21}(x, y)$ holds whenever the "true" predictor is asymmetrically distributed without imposing distributional assumptions on the error. Thus, one obtains $\rho^2_{12}(x, y) < \rho^2_{21}(x, y)$ under $x \rightarrow y$ and $\rho^2_{12}(x, y) > \rho^2_{21}(x, y)$ under $y \rightarrow x$ independent of the error term distribution. A nonparametric bootstrap approach can again be carried out for statistical inference. Similarly, kurtosis-based DDA measures can be obtained when focusing on $\rho^2_{13}(x, y)$ and $\rho^2_{31}(x, y)$ (cf. Wiedermann, 2017). Implementing additional DDA measures for potentially nonnormal "true" errors in `DDA_VarDist` is planned in the future.

Methods to assess independence of predictor(s) and error can straightforwardly be applied without any further modification even when the "true" error is nonnormal. The reason for this is that the Darmois–Skitovich theorem, as applied in the present context, does not impose distributional assumptions on the "true" error. Nonindependence of predictor(s) and error will hold when at least one common variable is nonnormal. Thus, evaluating the independence assumption of competing models can be carried out when (1) only the "true" predictor, (2) only the "true" error, or (3) both deviate from normality as along as the product of corresponding coefficients (see Eq. 18) is unequal to zero. However, results of competing BP-tests to assess patterns of heteroscedasticity in the two candidate models must be interpreted with caution when residuals of both models deviate from normality. In this case, Type I error rates of the test will be distorted and directional decisions must be based on Koenker's robust BP test.

Power and sample size considerations: What we know so far

To provide guidelines for the necessary number of observations to achieve sufficient power, we summarize previous simulation studies on DDA components and focus on three factors that impact empirical power rates: The magnitude of nonnormality, the magnitude of the causal effects, and sample size. Dodge and Rousson (2016) evaluated the power of nonparametric bootstrap CIs of $\Delta(\gamma) = |\gamma_x|-|\gamma_y|$ and $\Delta(\kappa) = \kappa_x|-|\kappa_y|$ and concluded that skewness-based model selection outperformed the kurtosis-based approach in terms of statistical power to detect the "true" model. Here, for small effects ($R^2 = .25$) and skewness values of 2, sample sizes as small as $n = 50$ may be sufficient to achieve a statistical power close to 80%. In contrast, for kurtosis-based selection, sample sizes of $n = 500$ and excess-kurtosis values larger 4 are needed to achieve similar statistical power.

Wiedermann and von Eye (2015b) evaluated power properties of residual distribution-based methods considering separate D'Agostino skewness tests, the asymptotic skewness difference test, and nonparametric bootstrap CIs for $\Delta(\gamma_e) = |\gamma_{e_{xy}}|-|\gamma_{e_{yx}}|$, and concluded that acceptable power levels can already be observed for $n = 75$ when causal effects are small ($\rho_{xy} = .25$) and the true predictor is sufficiently skewed (i.e., $\gamma_x \geq 2$). Because model selection based on separate normality tests proves more powerful than tests based on $\Delta(\gamma_e)$, $n = 50$ may already be sufficient for separate D'Agostino tests. In general, at least $n = 125$ is required for less skewed variables (e.g., $\gamma_x = 1$) and lower correlations (e.g., $\rho_{xy} = .25$). Model selection based on excess-kurtosis differences of residual distributions was evaluated by Wiedermann (2015). Again, separate Anscombe–Glynn tests outperformed procedures based on the difference of excess-kurtosis estimates. Here, for $n = 200$ and $\rho_{xy} = .4$, excess-kurtosis values larger than 4 are necessary for power rates close to 80%.

Wiedermann, Artner, and von Eye (2017) compared the performance of nine homoscedasticity tests to evaluate the independence assumption in competing models and showed that the BP-test was the most powerful procedure to select the correct model. For slightly skewed predictors ($\gamma_x = 0.75$), large effects and large sample sizes $n \geq 400$ may be required to achieve sufficient power. For $\gamma_x \geq 1.5$ and medium effect sizes, at least $n = 200$ may be required. Quite similar results were obtained for model selection based on nonlinear correlation tests of the form $cor\left(x, e^2_{yx}\right)$ (Wiedermann & von Eye, 2016). However, $\gamma_x \geq 1.5$ and large effects are necessary to obtain power rates beyond 80% when $n \geq 200$. Systematic simulation experiments that (1) compare the statistical power of several other independence tests and (2) evaluate all DDA components simultaneously constitute important future endeavors.

Further application scenarios and extensions

It is important to note that the proposed method is not restricted to the presented standard multiple regression setup. DDA is also applicable in other scenarios in which directionality issues have been deemed to be untestable. For example, when a statistical relation between two variables, x and y, has been established, researcher may further entertain hypotheses about the role of a third measured variable. Whether this third variable (m) should be conceptualized as a mediator (an intervening variable that transmits the effect from x to y) or as an observed confounder cannot be answered with standard statistical methods (MacKinnon, Krull, & Lockwood, 2000). From a DDA perspective, distinguishing between these models reduces to separately evaluating the directionality of x and m (i.e., whether $x \rightarrow m$ or $m \rightarrow x$ should be preferred) and m and y (i.e., whether $m \rightarrow y$ or $y \rightarrow m$ holds for the data) provided that nonnormality requirements are fulfilled (for extensions of residual- and independence-based DDA to mediation models, see Wiedermann & von Eye, 2015c, 2016). Furthermore, the application of DDA may not be restricted to observational studies. Directionality issues may also occur in experimental

studies—in particular, those designed to test hypotheses that go beyond total effects in randomized trials. Here, mediation models may, again, provide sound explanations how experimental interventions causally affect the target outcome (Bullock, Green, & Ha, 2010; Heckman & Smith, 1995; Imai, Keele, & Tingley, 2010). However, even when the predictor is under experimental control, it is well-known that neither the direction (Wiedermann & von Eye, 2015c) nor the magnitude of the causal effect of the mediator on the outcome can be identified uniquely without imposing strong assumptions on data, assumptions that are similar to observational studies (Imai, Tingley, & Yamamoto, 2013; Keele, 2015). Again, DDA may help to gain further insight through evaluating competing mediator-outcome paths while adjusting for an experimentally controlled predictor.

Extensions of the direction dependence methodology proposed in this article can go in a number of directions. First, developing DDA for moderation models would enable researchers to test the direction of effect while accounting for a third variable that modifies the relation between predictor and response (the fact that the nature of the moderator effect may depend on the direction of the postulated model has been shown by Judd & Kenny, 2010). Similarly, future work is needed to study principles of direction dependence in polynomial (i.e., models that consider higher-order terms, cf. Aiken & West, 1991) and more general linearizable regression models (i.e., nonlinear regression functions that can be linearized through proper variable transformations). Another possible extension concerns the complexity of the research design. Although the presented framework is designed for single-level data, developing DDA for multilevel regression models (Raudenbush & Bryk, 2002) would allow to account for hierarchical (nested) data structures. Further, throughout the article, we assumed that the "true" predictor is measured without measurement error. Although first attempts to extend DDA components to measurement error models are given in von Eye and Wiedermann (2014) and Wiedermann, Merkle, and von Eye (2018), extending direction dependence to latent variable models may overcome potential biases in directional decisions resulting from imprecise measurement of constructs. Finally, the present study focused on cases in which the tentative predictor and the tentative response are continuous variables (covariates can either be continuous or categorical). The reason for this is that both candidates models ($x \rightarrow y$ and $y \rightarrow x$) must be specified as standard linear regression models (similarly, the proposed SPSS macros are designed to evaluate two competing standard linear models). Although previous studies (cf. Inazumi et al., 2011; Peters, Janzing, & Schölkopf, 2011; von Eye & Wiedermann, 2016, 2017; Wiedermann & von Eye, 2018) discussed principles of direction dependence when both variables are categorical in nature, extending DDA to the generalized linear modeling framework (McCullagh & Nelder, 1989) would be most promising for evaluating causal relations among categorical, count, and continuous variables.

Author note We thank the two anonymous reviewers, Wes Bonifay, Francis Huang, Edgar C. Merkle, Anna P. Nutt, and Phillip K. Wood for their constructive comments on an earlier version of the article. We are also indebted to Ingrid Koller for providing the data used for illustrative purposes.

References

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Thousand Oaks: Sage.

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

Anscombe, F. J., & Glynn, W. J. (1983). Distribution of the kurtosis statistics b2 for normal samples. *Biometrika*, *70*, 227–234. doi:https://doi.org/10.2307/2335960

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. New York: Wiley.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology*, *9*, 78–84. doi:https://doi.org/10.1027/1614-2241/a000057

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Box, G. E. P., & Watson, G. S. (1962). Robustness to nonnormality of regression tests. *Biometrika*, *49*, 93–106. doi:https://doi.org/10.1093/biomet/49.1-2.93

Braun, M. T., & Oswald, F. L. (2011). Exploratory regression analysis: A tool for selecting models and determining predictor importance. *Behavior Research Methods*, *43*, 331–339. doi:https://doi.org/10.3758/s13428-010-0046-8

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, *98*, 550–558. doi:https://doi.org/10.1037/a0018933

Cain, M. K., Zhang, Z., & Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716–1735. doi:https://doi.org/10.3758/s13428-016-0814-1

Chickering D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507–554.

Cook, D. L. (1959). A replication of Lord's study on skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement*, *19*, 81–87. doi:https://doi.org/10.1177/001316445901900109

Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman & Hall.

Cudeck, R., & Henly, S. J. (2003). A realistic perspective on pattern representation in growth data: Comment on Bauer and Curran (2003). *Psychological Methods*, *8*, 378–383. doi:https://doi.org/10.1037/1082-989X.8.3.378

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large sample sizes. *Biometrika*, *58*, 341–348. doi:https://doi.org/10.2307/2334522

Darmois, G. (1953). Analyse générale des liaisons stochastique. *Review of the International Statistical Institute*, *21*, 2–8. doi:https://doi.org/10.2307/1401511

Dehaene, S., & Cohen, L. (1998). Levels of representation in number processing. In B. Stemmer & H. A. Whitaker (Eds.), The handbook of neurolinguistics (pp. 331–341). New York: Academic Press.

Dodge, Y., & Rousson, V. (2000). Direction dependence in a regression line. *Communications in Statistics: Theory and Methods*, *29*, 1957–1972. doi:https://doi.org/10.1080/03610920008832589

Dodge, Y., & Rousson, V. (2001). On asymmetric properties of the correlation coefficient in the regression setting. *American Statistician*, *55*, 51–54. doi:https://doi.org/10.1198/000313001300339932

Dodge, Y., & Rousson, V. (2016). Recent developments on the direction of a regression line. In W. Wiedermann & A. von Eye (eds.), Statistics and causality: Methods for applied empirical research (pp. 45–62). Hoboken: Wiley.

Dodge, Y., & Yadegari, I. (2010). On direction of dependence. *Metrika*, *72*, 139–150. doi:https://doi.org/10.1007/s00184-009-0273-0

Entner, D., Hoyer, P. O., & Spirtes, P. (2012). Statistical test for consistent estimation of causal effects in linear non-Gaussian models. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, *22*, 364–372.

Frisch, R., & Waugh, F. (1933). Partial time regressions as compared with individual trends. *Econometrica*, *1*, 387–401. doi:https://doi.org/10.2307/1907330

Geisser, J. (1993). Predictive inference: An introduction. London: Chapman & Hall.

Gentile, D. A., Lynch, P. J., Linder, J. R., & Walsh, D. A. (2004). The effects of violent video game habits on adolescent hostility, aggressive behaviors, and school performance. *Journal of Adolescence*, *27*, 5–22. doi:https://doi.org/10.1016/j.adolescence.2003.10.002

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2008). A kernel statistical test of independence. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), Advances in neural information processing systems (Vol. 20, pp. 585–592). Cambridge: MIT Press.

Hampel, F. R. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie*, *27*, 87–104. doi:https://doi.org/10.1007/bf00536619

Harris, A., & Seckl, J. (2011). Glucocorticoids, prenatal stress and the programming of disease. *Hormones and Behavior*, *59*, 279–289. doi:https://doi.org/10.1016/j.yhbeh.2010.06.007

Heckman, J. J., & Smith, J. A. (1995) Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*, 85–110. doi:https://doi.org/10.1257/jep.9.2.85

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*, 365–388. doi:https://doi.org/10.1177/0013164414548576

Hoyer, P. O., Shimizu, S., Kerminen, A. J., & Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, *49*, 362–378. doi:https://doi.org/10.1016/j.ijar.2008.02.006

Hyvärinen, A. (2010). Pairwise measures of causal direction in linear non-Gaussian acyclic models. In *JMLR: Workshop and Conference Proceedings* (Vol. 13, pp. 1–16). Tokyo, Japan: JMLR.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). Independent components analysis. New York: Wiley.

Hyvärinen, A., & Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *Journal of Machine Learning Research*, *14*, 111–152.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309–334. doi:https://doi.org/10.1037/a0020761

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A*, *176*, 5–51. doi:https://doi.org/10.1111/j.1467-985x.2012.01032.x

Inazumi, T., Washio, T., Shimizu, S., Suzuki, J., Yamamoto, A., & Kawahara, Y. (2011). Discovering causal structures in binary exclusive-or skew acyclic models. In F. Cozman & A. Pfeffer (Eds.), Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (pp. 373–382). Corvallis: AUAI Press. arXiv:1202.3736

James, L. R., & Singh, B. K. (1978). An introduction to the logic, assumptions, and basic analytic procedures of two-stage least squares. *Psychological Bulletin*, *85*, 1104–1122. doi:10.1037/0033-2909.85.5.1104

Judd, C. M., & Kenny, D. A. (2010). Data analysis. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (5th ed., Vol. 1, pp. 115–139). New York: Wiley.

Kaufman, R. L. (2013). Heteroskedasticity in regression: Detection and correction. Thousand Oaks: Sage.

Keele, L. (2015). Causal mediation analysis: Warning! Assumptions ahead. *American Journal of Evaluation*, *36*, 500–513. doi:https://doi.org/10.1177/1098214015594689

Koller, I., & Alexandrowicz, R. W. (2010). A psychometric analysis of the ZAREKI-R using Rasch-models. *Diagnostica*, *56*, 57–67. doi:https://doi.org/10.1026/0012-1924/a000003

Lim, C. R., Harris, K., Dawson, J., Beard, D. J., Fitzpatrick, R., & Price, A. J. (2015). Floor and ceiling effects in the OHS: An analysis of the NHS PROMs data set. *BMJ Open*, *5*, e007765. doi:https://doi.org/10.1136/bmjopen-2015-007765

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement*, *15*, 383–389. doi:https://doi.org/10.1177/001316445501500406

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading: Addison-Wesley.

Lovell, M. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, *58*, 993–1010. doi:https://doi.org/10.1080/01621459.1963.10480682.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*, 173–181. doi:https://doi.org/10.1023/A:1026595011371

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:https://doi.org/10.1007/bf02296272

McCullagh, P., & Nelder, A. (1989). Generalized linear models (2nd). London: Chapman & Hall.

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

Miller, T. W., Nigg, J. T., & Miller, R. L. (2009). Attention deficit hyperactivity disorder in African American children: What can be concluded from the past ten years? *Clinical Psychology Review*, *29*, 77–86. doi:https://doi.org/10.1016/j.cpr.2008.10.001

Muddapur, M. V. (2003). On directional dependence in a regression line. *Communications in Statistics: Theory and Methods*, *32*, 2053–2057. doi:https://doi.org/10.1081/sta-120023266

Mumford, J. A., & Ramsey, J. D. (2014). Bayesian networks for fMRI: A primer. *NeuroImage*, *86*, 573–582. doi:https://doi.org/10.1016/j.neuroimage.2013.10.020

Munafò, M. R., & Araya, R. (2010). Cigarette smoking and depression: A question of causation. *British Journal of Psychiatry*, *196*, 425–426. doi:https://doi.org/10.1192/bjp.bp.109.074880

Nigg, J. T. (2012). Future directions in ADHD etiology research. *Journal of Clinical Child & Adolescent Psychology*, *41*, 524–533. doi:https://doi.org/10.1080/15374416.2012.686870

Nigg, J. T., Knottnerus, G. M., Martel, M. M., Nikolas, M., Cavanagh, K., Karmaus, W., & Rappley, M. D. (2008). Low blood lead levels associated with clinically diagnosed attention-deficit/hyperactivity disorder and mediated by weak cognitive control. *Biological*

Psychiatry, *63*, 325–331. doi:https://doi.org/10.1016/j.biopsych.2007.07.013

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd). Cambridge: Cambridge University Press.

Pearson, E. S. (1931). The analysis of variance in case of non-normal variation. *Biometrika*, *23*, 114–133. doi:https://doi.org/10.2307/2333631

Peters, J., Janzing, D., & Schölkopf, B. (2011). Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*, 2436–2450. doi:https://doi.org/10.1109/tpami.2011.71

Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: Foundations and learning algorithms. Cambridge: MIT Press.

Pornprasertmanit, S., & Little, T. D. (2012). Determining directional dependency in causal associations. *International Journal of Behavioral Development*, *36*, 313–322. doi:https://doi.org/10.1177/0165025412448944

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press. (Original work published 1960)

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks: Sage.

Richardson, T., & Spirtes, P., (1999). Automated discovery of linear feedback models. In C. Glymour & G. F. Cooper (Eds.), Computation, causation and discovery (pp. 253–304). Cambridge: MIT Press.

Rogosa, D. R. (1985). Analysis of reciprocal effects. In T. Husen & N. Postlethwaite (Eds.), International encyclopedia of education (pp. 4221–4225). London: Pergamon Press.

Sen, A., & Sen, B. (2014). Testing independence and goodness-of-fit in linear models. *Biometrika*, *101*, 927–942. doi:https://doi.org/10.1093/biomet/asu026

Shimizu, S. (2016). Non-Gaussian structural equation models for causal discovery. In W. Wiedermann & A. von Eye (eds.), Statistics and causality: Methods for applied empirical research (pp. 153–276). Hoboken: Wiley.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. J. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., . . . Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, *12*, 1225–1248.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310. doi:https://doi.org/10.1214/10-sts330

Skitovich, W. P. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, *89*, 217–219.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search (2nd). Cambridge: MIT Press

Spirtes, P., Richardson, T., Meek, C., Scheines, R., & Glymour, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociological Methods and Research*, *27*, 182–225. doi:https://doi.org/10.1177/0049124198027002003

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, *3*, 1–28. doi:https://doi.org/10.1186/s40535-016-0018-x

Sungur, E. A. (2005). A note on directional dependence in regression setting. *Communications in Statistics: Theory and Methods*, *34*, 1957–1965. doi:https://doi.org/10.1080/03610920500201228

Taylor, G., McNeill, A., Girling, A., Farley, A., Lindson-Hawley, N., & Aveyard, P. (2014). Change in mental health after smoking cessation: Systematic review and meta-analysis. *British Medical Journal*, *348*, 1–22. doi:https://doi.org/10.1136/bmj.g1151

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., … de Vet, H. C. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*, 34–42. doi:https://doi.org/10.1016/j.jclinepi.2006.03.012

Teuscher, F., & Guiard, V. (1995). Sharp inequalities between skewness and kurtosis for unimodal distributions. *Statistics and Probability Letters*, *22*, 257–260. doi:https://doi.org/10.1016/0167715294000074I

Verma, T. S., & Pearl, J. (1991). Equivalence and synthesis of causal models. *Uncertainty in Artificial Intelligence*, *6*, 220–227.

von Aster, M., Weinhold Zulauf, M., & Horn, R. (2006). Neuropsychologische Testbatterie fuer Zahlenverarbeitung und Rechnen bei Kindern (ZAREKI-R) [Neuropsychological test battery for number processing and calculation in children]. Frankfurt: Harcourt Test Services.

von Aster, M. G., & Shalev, R. S. (2007). Number development and dyscalculia. *Developmental Medicine and Child Neurology*, *49*, 868–873. doi:https://doi.org/10.1111/j.1469-8749.2007.00868.x

von Eye, A., & DeShon, R. P. (2012). Directional dependence in developmental research. *International Journal of Behavioral Development*, *36*, 303–312. doi:https://doi.org/10.1177/0165025412439968

von Eye, A., & Wiedermann, W. (2014). On direction of dependence in latent variable contexts. Educational and Psychological Measurement, 74(1), 5–30. doi:https://doi.org/10.1177/0013164413505863

von Eye, A., & Wiedermann, W. (2016). Direction of effects in categorical variables: A structural perspective. In W. Wiedermann & A. von Eye (Eds.), Statistics and causality: Methods for applied empirical research (pp. 107–130). Hoboken: Wiley.

von Eye, A., & Wiedermann, W. (2017). Direction of effects in categorical variables: Looking inside the table. *Journal of Person-Oriented Research*, *3*, 11–26. doi:https://doi.org/10.17505/jpor.2017.02

White, H., & MacDonald, G. M. (1980). Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association*, *75*, 16–28. doi:https://doi.org/10.2307/2287373

Wiedermann, W. (2015). Decisions concerning the direction of effects in linear regression models using the fourth central moment. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), Dependent data in social sciences research: Forms, issues, and methods of analysis (pp. 149–169). New York: Springer.

Wiedermann, W. (2017). A note on fourth moment-based direction dependence measures when regression errors are non normal. *Communications in Statistics: Theory and Methods*. doi:https://doi.org/10.1080/03610926.2017.1388403

Wiedermann, W., Artner, R., & von Eye, A. (2017). Heteroscedasticity as a basis of direction dependence in reversible linear regression models. *Multivariate Behavioral Research*, *52*, 222–241. doi:https://doi.org/10.1080/00273171.2016.1275498

Wiedermann, W., & Hagmann, M. (2015). Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals. *Communications in Statistics*, *45*, 6263–6283. doi:https://doi.org/10.1080/03610926.2014.960582

Wiedermann, W., Hagmann, M., Kossmeier, M., & von Eye, A. (2013). Resampling techniques to determine direction of effects in linear regression models. *Interstat*. Retrieved May 13, 2013, from http://interstat.statjournals.net/YEAR/2013/articles/1305002.pdf

Wiedermann, W., Hagmann, M., & von Eye, A. (2015). Significance tests to determine the direction of effects in linear regression models. *British Journal of Mathematical and Statistical Psychology*, *68*, 116–141. doi:https://doi.org/10.1111/bmsp.12037

Wiedermann, W., Merkle, E. C., & von Eye, A. (2018). Direction of dependence in measurement error models. *British Journal of Mathematical and Statistical Psychology*, *71*, 117–145. doi:https://doi.org/10.1111/bmsp.12111

Wiedermann, W., & von Eye, A. (2015a). Direction-dependence analysis: A confirmatory approach for testing directional theories.

International Journal of Behavioral Development, *39*, 570–580. doi:https://doi.org/10.1177/0165025415582056

Wiedermann, W., & von Eye, A. (2015b). Direction of effects in multiple linear regression model. *Multivariate Behavioral Research*, *50*, 23–40. doi:https://doi.org/10.1080/00273171.2014.958429

Wiedermann, W., & von Eye, A. (2015c). Direction of effects in mediation analysis. *Psychological Methods*, *20*, 221–244. doi:https://doi.org/10.1037/met0000027

Wiedermann, W., & von Eye, A. (2016). Directionality of effects in causal mediation analysis. In W. Wiedermann & A. von Eye (Eds.), Statistics and causality: Methods for applied empirical research (pp. 63–106). Hoboken: Wiley.

Wiedermann, W., & von Eye, A. (2018). Log-linear models to evaluate direction of effect in binary variables. *Statistical Papers*. doi:https://doi.org/10.1007/s00362-017-0936-2

Wong, C. S., & Law, K. S. (1999). Testing reciprocal relations by nonrecursive structural equation models using cross-sectional data. *Organizational Research Methods*, *2*, 69–87. doi:https://doi.org/10.1177/109442819921005

Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, *172*, 1873–1896. doi:https://doi.org/10.1016/j.artint.2008.08.001