PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2023, TOM 66, NR 1, 209–240
DOI: 10.31648/przegldpsychologiczny.9731

Introduction to Causal Inference for Psychologists: Testable and Non-Testable Causal and Statistical Assumptions

Borysław Paulewicz[1]
Jagiellonian University, Institute of Psychology
https://orcid.org/0000-0002-1270-2988

Abstract

The main goal of basic research is to answer causal questions. Generally, only the statistical part of this process tends to proceed in a partially formal way and according to clearly defined rules. At the same time, causal relations are often treated informally or implicitly in a way that is prone to difficult-to-detect errors. This introduction aims to show psychology researchers some of the great benefits of approaching causal issues using a formal theory of causal inference. In this part, I discuss the non-obvious status and role of causal and statistical assumptions in causal inference. After covering, in a simple setting, the general shape of inference from causal assumptions, statistical assumptions, and data to causal effects, I outline, from a contemporary perspective, the limits of applicability of the general linear model. Then, I introduce the formal part of Pearl's theory that relies on graphs. Using these tools, I show how one can analyze and interpret the results of an experiment on short-term memory search, and I discuss the back-door and front-door adjustments. To present the mathematical part of the theory in an accessible way without overly simplifying it, I illustrate some issues by using simulations written in R.

Keywords: causality, · causal inference, causal calculus, research methods, metatheory, statistical inference, Bayesian inference

The main goal of this introduction is to convince psychologists who conduct or rely on the results of scientific research that they should not ignore causal inference theory. This is because its theorems, deduced from axioms expressing

[1] Correspondence address: boryslaw.paulewicz@uj.edu.pl. Comments and corrections: https://github.com/boryspaulewicz/przyczynowosc1.

elementary intuitions about causality, are of great importance to the entire field of psychological methods.

I have tried to make this introduction accessible to readers with little mathematical training. However, I could not entirely avoid mathematics because the theory of causal inference that I describe here is mathematical, which makes the level of difficulty of what follows quite variable. My experiences in teaching the basics led me to appreciate the didactic usefulness of computer simulations. Nowadays, many psychologists can create and alter such simulations, and doing so facilitates familiarity with the essential concepts before mastering the technical definitions, which, for many, does not come quickly. This should also make it easier for readers not accustomed to manipulating expressions denoting probability distributions to gain some understanding of the mathematical part of this text and, consequently, to engage with the literature on the subject, where such expressions are common.

I rely on Pearl's Structural Causal Model (SCM, see Pearl, 2000; Pearl et al., 2016; Pearl & Mackenzie, 2021) and the associated calculus, i.e., the do-calculus. The most important alternative theory of similar status is Neyman and Rubin's Potential Outcomes framework (Rubin, 2005), which I do not discuss because the axioms of one theory follow from the axioms of the other, and vice versa (Galles & Pearl, 1998). Not only does Pearl's version seem to be under more intensive development and it is arguably easier to use, but also – and more importantly for my purposes – it seems to be more readily applicable in psychology thanks to the fact that not everything in it needs to be defined using counterfactual notions.

The way I explain the basics is, I hope, mostly standard. Still, it differs somewhat from how Pearl's theory is presented in the popular science "The Book of Why" (Pearl & Mackenzie, 2021), in "Causal Inference in Statistics: A Primer" (Pearl et al., 2016), which is a comprehensive textbook for beginners or intermediate learners, or in the relatively challenging "Causality" (Pearl, 2000). I try to be consistently explicit about the fact that, by default, causal relations represented by graphs indicate theoretical possibilities. I repeatedly emphasize that, typically, most causal assumptions are not testable. I dedicate some space to the role of statistical assumptions and the statistical methods most commonly used by psychologists. I discuss examples of psychological research, including one in some detail. Last but not least, I also refer to the two most popular Polish handbooks on statistics and research methods written for psychologists. To minimize the risk of misunderstanding, I liberally use emphasis by writing words or phrases using italics. I would like to think that thanks to all this, what follows may be tailored to the needs of the intended readers.

"Correlation Is Not Causation"

Let's begin by considering the correlation between yearly income (X) and life satisfaction (Y). Unlike self-reported life satisfaction, *actual* life satisfaction is not directly observable, but at this point, for simplicity, I will ignore this problem.

We cannot demand that non-trivial empirical statements that appear as steps in scientific reasoning or as assumptions of scientific theories be certainly true since empirical statements are not mathematical statements. However, we should typically require such statements to be empirically or theoretically *justified*. Claiming that there is a direct causal effect based on correlation *alone* is entirely unjustified. It's not surprising then that researchers widely view this mistake as serious as well as elementary. And yet, a closer analysis shows that it is far from obvious precisely when and in what sense correlation does not imply causation.

In the ongoing example, we have two types of assumptions. Some are related to the design of the study:

1.1 X is observed.

1.2 Y is observed.

… and some are related to statistical analysis and its results:

1.3 X and Y are positively correlated.

There is also the candidate conclusion:

1.4 X has a causal effect on Y.

Given only premises 1.1–1.3, the conclusion (1.4) that X influences Y is not justified, although this does not mean it is untrue. The observed statistical dependence (here, a correlation) may result, perhaps entirely, from the causal effect of some third variable on X and Y, or from the "reverse" influence of Y on X; assumptions 1.1–1.3 do not exclude these two alternative causal explanations. Suppose we introduced some theoretical or empirical arguments that are independent of the study and its results and that justify the causal conclusion. In that case, the conclusion would be justified based on these independent arguments, but it would still not be justified at all *as a conclusion of the study*. From the perspective of a critical reader, such a conclusion as a conclusion of the study would be pulled out of thin air.

Statistical inference is applied probability calculus, which is a theory of distributions, i.e., of relative frequencies of or subjective confidence in the occurrence of possible events or outcomes when the data-generating process is *fixed*. When we talk about causation, we must consider, within the same model, *different possible data-generating processes*, not just the probabilities of various events. Therefore, just as we cannot justify statistical claims without first introducing statistical assumptions, justifying causal claims requires introducing causal assumptions.

As far as causality is concerned, in this case it is important that X is observed but not randomly assigned. However, this property cannot be expressed in the language of statistics because it has nothing to do with how often the possible values of X may occur. Such assumptions about the properties of study design are important because they may help eliminate alternative causal explanations of the observed statistical effects. The minimal logically correct version of the reasoning under consideration must then look something like this:

2.1 Y does not have a causal effect on X.

2.2 X and Y do not have common causes.

2.3 and Y are correlated.

Therefore:

2.4 X has a causal effect on Y.

Assumption 2.3 is the only statistical assumption (it is also a conclusion from the data and the omitted statistical assumptions), while assumptions 2.1–2.2 are causal. Assumptions 2.1 and 2.2 will be justified when X is randomly assigned, for instance. In such a case, no variable other than the randomization device or process will have a causal effect on X, and this will make assumptions 2.1 and 2.2 true.

In reasoning 2.1–2.4, we implicitly rely on the meta-assumption to the effect that statistical dependence, unless it is merely a result of sampling error, has to have some causal or structural origin. In this sense, causation does follow from a nonzero correlation, but correlation alone is not enough to determine the direction of causality. Even when the observed correlation is an artifact, this artifact must also arise from the causal structure of the data-generating process. That is why the assumption that X *may* influence Y does not have to appear in the correct version of the reasoning; introducing this quasi-assumption would perhaps increase readability, but, given the meta-assumption, it would be unnecessary. That is, as long as the expression “may cause” means that it may or may not, i.e., that we do not make a claim either way.

The meta-assumption regarding the nonexistence of “accidental” statistical dependencies in the population can perhaps be replaced by a weaker version like “a statistical dependence in the population may arise from invariant causal relations between variables, or it may be accidental.”[2] In practice, it doesn’t seem to matter, as we can never be certain that a statistical effect is accidental in this sense.

Let’s then imagine that X was randomized. In that case, the assumptions regarding the absence of a direct causal effect of Y on X and the absence of any common causes would be justified based on the assumption of randomization, which is a known property of study design. Simultaneously, none of the justified causal assumptions, which are precisely assumptions regarding the absence of certain causal effects, would be *testable* in this situation. Even in this simple example, we can then see that a causal model can be simultaneously justified, useful, and non-testable based on the results of the study it’s meant to describe.

Some Remarks on Linear Regression

One of my goals when writing this section was to clarify certain issues related to statistical models commonly used in psychology that a reader who may

[2] The definition of statistical dependence does not require it to result from any causal relation. Therefore, statistical dependence can be accidental in at least two senses: due to sampling error, it may only appear that the dependence exists in the distribution or the population, or the dependence may actually occur in the population but only due to some random coincidence, e.g., because the population is finite and it just so happens that the values of a particular variable do not have the same distribution for every value of some other variable(s). We can be faced with this latter possibility, however, only if we are concerned with a particular population at a particular moment in time rather than an invariant, abstract data-generating process.

have learned about them from "Statystyczny drogowskaz" [A statistical signpost] (Bedyńska et al., 2012) or "Metodologia badań psychologicznych" [Methodology of psychology research] (Brzeziński, 2022), might be confused about. Since our first example involves correlation, the conclusion relies in part on the use of *linear regression*, which is a statistical model that has the following general form:

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_n X_n + \epsilon$$

where Y is the *dependent variable*, X_i are *predictors* or *independent variables*, α_i are *regression coefficients*, and ϵ is the *regression error*, which is assumed to have a normal distribution, with mean 0 and standard deviation σ. Greek letters represent *free parameters*, i.e., quantities that are usually unknown but can, in principle, be estimated from data. The expression on the right-hand side, excluding ϵ, is the *systematic part* describing the expected value (or mean) of the distribution of Y as a function of the predictors.

I'm using contemporary terminology, according to which a regression model – whether linear or not – is any statistical model describing a *conditional distribution*. In this case, it is a model of the conditional distribution of Y. This distribution, which we can denote concisely as $p(Y \mid X_1,\ldots,X_n)$, is defined as a function of the variables X_i, and as far as this distribution is concerned, all the X_is are treated as known constants. ANOVA[3] models are then also regression models because they describe (a property of) the distribution of the dependent variable as a function of some (nominal) variables.

The correlation coefficient mentioned earlier is equal to the slope of the regression line if both the (only) predictor and the dependent variable have standard deviations of 1, which can be arranged by dividing the two variables by their respective standard deviations. If we subtract from the observed values of Y the values implied by the systematic part of the fitted model, we obtain the *residuals*; these resemble the regression errors, but since the model is only estimated – not known – they are not the same as regression errors.

Contrary to what the authors of the two mentioned handbooks seem to at least suggest, this regression is linear not because it can describe only straight lines or planes but because it is *linear in the parameters*. For example, the model $Y = X^{\alpha} + \epsilon$ is nonlinear (in the parameters), but $Y = \alpha \sin^2(X) + \epsilon$, where $\sin^2(X)$ is just a transformed independent variable, is simultaneously curvilinear (with respect to X) and linear (in the parameters) because by denoting $\sin^2(X)$ as Z, we can equivalently represent this latter model as $Y = \alpha_0 + \alpha_1 Z_1 + \epsilon$, thus meeting the requirements of the definition of linear regression.

In psychological research, the assumptions of linear regression are never jointly true because they cannot possibly be. The observable variables studied by psychologists are typically discrete and bounded from below and from above.

[3] Brzeziński (2022) maintains that ANOVA models are suitable for analyzing results from experimental studies, while regression models, in his opinion, are suitable for analyzing results from observational studies, which he refers to as studies "corresponding to the correlational model."

Even when an observed variable of interest to psychology is continuous, it is at least bounded from one side, e.g., reaction time is always positive. Regression errors cannot then have a normal distribution because every normal distribution is continuous and unbounded. No transformation will guarantee that the distribution of regression errors is normal even when the dependent variable is continuous. For such a transformation to be known, the actual distribution of the residuals would also have to be known, but we do not know much about the distributions of the observable continuous variables studied in psychology. Hence, testing if the distribution of residuals deviates from normality, which is a practice recommended in the two aforementioned handbooks, does not make much sense. This kind of test is not only useless as a normality test in psychological research, but it is also not a good indicator of the extent to which violation of the normality assumption justifies using alternative methods. The interested reader can learn more about all this from contemporary handbooks on robust statistical methods (e.g., Wilcox, 2011), i.e., methods that were designed to work under conditions where statistical model assumptions are not true. Such methods are modern alternatives to the classical nonparametric methods that are recommended by the authors of the two aforementioned handbooks and that were invented before any theory of robust statistics was proposed.

Point estimates of regression coefficients are often obtained using the method of ordinary least squares (OLS). Importantly, *OLS estimates of linear regression coefficients have a descriptive sense even when the assumptions of linear regression are* f a l s e. This universal descriptive usefulness reflects the fact that OLS estimates of linear regression coefficients always correspond to the orthogonal projection of the dependent variable vector onto the linear plane spanned by the predictor vectors[4]. For instance, the OLS sample estimate of the intercept and slope is an unbiased estimate of *the line that minimizes the mean square error*. This is true regardless of whether the error distribution is normal (which it never is in psychology) and has constant variance (which it probably doesn't) or whether the data are independent given the model (often, it cannot be known that they are; see, e.g., Greenland, 2022), or whether the systematic part of the true relationship is best described by a linear equation (it probably isn't) or the linear (in the parameters) effects exist only "virtually" as "trends." Therefore, one can always say that correlation represents the degree of *linear* covariation, although it is rarely, if ever, correct to say that correlation represents the degree of covariation. Similarly, the sample mean is *always* an unbiased estimate of the mean of the distribution from which the sample was drawn because of the linearity of the expected value operator. Lack of bias means that the estimates obtained in a hypothetical infinite series of replications of the same sampling process should be equal, on average, to what is supposed to be estimated. So, in a sense, an unbiased estimate estimates what it is supposed to.

[4] If we also consider as a predictor the constant vector corresponding to the intercept term when this term is present in the model formula. An excellent introduction to linear models viewed geometrically can be found in Saville and Wood (2012).

The Central Limit Theorem implies that, in a wide range of situations, the distribution of a weighted sum of random variables that do not necessarily have identical distributions will approach a normal distribution in the limit as the number of summed variables increases. OLS point estimates of linear regression coefficients are weighted sums of random variables. It follows that, regardless of whether the linear regression model is true, *interval* estimates (confidence intervals) and the corresponding significance levels will often be, at worst, more conservative (wider) than can be achieved using other methods. Since we almost always know in psychology that the statistical model is not true, interval estimates can be interpreted only as approximate measures of uncertainty about the properties of a simplified or partial description of the modeled distribution[5]. In contrast to the often catastrophic consequences of false causal assumptions, such problems can be minimized by increasing the sample size or changing the analysis method.

One of the important reasons to care about the *degree and nature* of deviations from the assumptions of the chosen statistical model rather than *whether* these assumptions are true – as they generally cannot be, and even if they are, there is little we can do to make sure they are true – is the *efficiency of the estimator*, which is the degree to which point estimates vary (not necessarily around the true values!), on average, in hypothetical replications of the same sampling process. We want the variance of the point estimates to be as small as possible on average. Moreover, we want the estimates to vary around the true values, i.e., values corresponding to a *simplified but well-defined description of the modeled distribution*. Finally, we want the associated uncertainty to be as small as possible. The frequent occurrence of substantial deviations from the statistical model assumptions means that robust methods should work better than OLS linear regression estimates in psychology in both respects[6], as the latter kind of estimates are known to be, in a technical sense, the least robust (Field & Wilcox, 2017; Wilcox, 2011).

The choice of a statistical method is a complex issue, and it is impossible to provide a simple algorithm that identifies the best solution in typical situations. Therefore, readers who primarily use linear regression or some of its many generalizations may be comforted by the fact that, with appropriate precautions (which I cannot elaborate on due to space constraints), they can often rely on such models. A fitted, *false* statistical model may be *perfectly acceptable* in that it can be used to derive justified and correct conclusions as long as this model is interpreted as only *a part of a method for estimating a simplified or partial description of the true distribution*.

This also applies to most situations where the scale of the dependent variable might seem to disallow the use of a linear model. In the frequentist interpretation, when we say that outcomes of a certain kind have a distribution,

[5] The second edition of "Statistical Rethinking," covering both Bayesian inference and causal inference, is an excellent example of consistently applying this way of thinking about statistical models (McElreath, 2020).

[6] However, this usually comes at the cost of some bias.

we are assuming, perhaps implicitly, the existence of some theoretically repeatable sampling process. When the process is specified, certain conditions are met[7], and we assign numbers to the possible outcomes, then we bring into existence a distribution with a well-defined mean. For example, the mean of a random sample of values, coded 0 as or 1, of a *nominal* variable like gender at birth is an unbiased estimate of the population mean of numbers 0 and 1 when these are assigned to the gender at birth of every individual in the population. Despite the original variable being nominal, this mean has an obvious interpretation: it is just the proportion of women (or men) in the population. The use of a linear model to describe the distribution of a nominal variable with more than two values is inconvenient (it can be done, e.g., by encoding each possible value using a separate binary indicator variable), but such variables occur relatively rarely in psychology as dependent variables.

How much one should worry about the scale being ordinal is a matter of debate (see, e.g., Liddell & Kruschke, 2018; Paulewicz & Blaut, 2022). However, note that it is not possible to even articulate the "ordinal scale problem" without considering at least *two distinct variables*: one whose numerical values are supposed to say something about the origin of the ordinal outcomes, and the ordinal variable itself, usually also assuming that the first variable causes the second. For example, the same differences in places on the podium interpreted as numbers 1, 2, and 3 will not correspond to the same differences in race completion times, which are the causes of the places on the podium. However, these are two different variables; as long as we are interested in the *observed ordinal variable as such*, e.g., the response on a Likert scale expressed as an integer, and not in its *hypothetical unobserved source*, e.g., the actual degree of endorsement, the problem of scale does not arise, and OLS regression coefficients are unbiased estimates of well-defined – although psychologically less interesting – quantitative properties of a simplified description of the conditional distribution of such a variable.

The Causal Nature of Computer Simulations and the Notion of Intervention

It seems appropriate to introduce causal inference by discussing computer simulations because a computer is a programmable device, and programming is about *controlling how it operates*. In particular, through simulations, one can relatively easily illustrate – but not prove, as no simulation can serve as mathematical proof – the meaning and validity of causal inference theorems.

Running the following R code creates a process in which the only qualitative causal relation between X and Y is the causal effect of X on Y. For simplicity, all

[7] I added this qualification because not every distribution of numerical values has a well-defined mean. However, this issue is unrelated to the notion of measurement scale.

effects are linear, and all "completely random" variables (I will soon give them a different name) have a standard normal distribution.

```
set.seed(1234)
n = 10000
U_X = rnorm(n)
U_Y = rnorm(n)
X = U_X
Y = 1 + 2 * X + U_Y
```

The first two instructions serve as the set-up: the `set.seed(1234)` instruction sets the random seed to the arbitrarily chosen number 1234, making the simulation process reproducible, and `n` is just a name for the number of generated samples. The `rnorm(n)` instruction generates (pseudo)random samples from the standard (i.e., with mean `0` and variance of 1) normal distribution (`rnorm` is short for "random normal").

The equal sign here does not represent mathematical equality but rather the *assignment operation*. To evaluate this instruction, R first evaluates the expression on the right-hand side. For example, the text "10000" in the instruction `n = 10000` is interpreted as the number 10000, and a representation of this number is stored in memory. The obtained value is then assigned to the variable on the left-hand side. What appears on the equal sign's right-hand side is then the cause of the state or value of the variable on the left-hand side. The instruction `U_X = rnorm(n)` causes n pseudorandom samples from a standard normal distribution to become the values of the variable U_X. The following instruction `U_Y = rnorm(n)` results in the generation of *new and independent* samples from a standard normal distribution, which then become the values of U_Y. In the simulated process, Y does not directly or indirectly influence X because the value of the expression describing how X is generated depends neither directly nor indirectly on Y. It is also clear that X and Y do not have common causes.

In causal inference, variables that in a given context are thought of as arising (here pseudo-) randomly or in an unspecified way are called *exogenous*, meaning they originate from outside (of the model or the modeled process). Variables generated by the non-random part of the process are called *endogenous*, meaning they are generated inside (the model). By choosing what variables are endogenous, a researcher selects the part of reality that is subject to causal analysis, so this choice is arbitrary and usually results from what the researcher is interested in and what unobservable theoretical constructs the researcher believes may exist.

According to the convention sometimes used in the literature, endogenous variables are also called *modeled* variables, and they are denoted by capital letters, excluding the letter U; their corresponding exogenous sources are sometimes called *unmodeled* and are denoted by the letter U with an appropriate subscript. I will not henceforth refer to variables as exogenous or endogenous to emphasize that this distinction is not about the variables themselves but rather about how they are viewed. Because, for each modeled variable V, its corresponding unmodeled source U_V represents all unmodeled causes of V, if V has no modeled causes, like X here, then U_V represents *all* causes of V. Thus, the set of values of such a modeled variable can be identified without loss of generality with the set of values of its unmodeled source (here, $X = U_X$).

Because we control the data-generating process, we can easily *observe the effects of interventions* by replacing certain expressions with constants. For example, physically setting the value of X to 44 corresponds to the following version of the process:

```
set.seed(1234)
n = 10000
U_X = rnorm(n)
U_Y = rnorm(n)
X = 44
Y = 1 + 2 * X + U_Y
```

If W and V are modeled variables or sets of modeled variables, then the expression $p\big(W|do(V = v)\big)$ denotes the *interventional distribution* generated by the process arising from the original one by replacing the sources of variability in V with the constant (vector, if V is a set) v. For instance, the above code generates samples from the interventional distribution $p\big(Y|do(X = 44)\big)$. This distribution represents the so-called *total* causal effect of the intervention. By looking at the code, we can immediately see that an intervention of the form $do(Y = y)$ cannot change the distribution of X, i.e., $p\big(X|do(Y = y)\big) = p(X)$ for every y. That is what we mean when we say there is no causal effect Y of on X.

The formal language of do-calculus differs from the language of probability calculus only in the presence of the *do* operator. Using this abstract operator, we can also define interventions that cannot really be performed in practice, such as hypothetical interventions that directly affect *only* blood pressure or gender. This way, it is possible to formulate causal questions that cannot be directly answered by conducting a randomized study, yet it may still be possible to obtain justified answers to such questions using do-calculus when conditions permit.

Structure and Interpretation of Causal Graphs

Drawing a causal graph is equivalent to only saying which modeled variables do not directly cause which other modeled variables and which pairs of modeled variables do not have any unmodeled common causes. Regardless of study design and despite the inherent difficulties of studying unobservable psychological response processes, namely the difficulties arising from the hard-to-predict, multidimensional intra- and interindividual variability of such processes, we can then *always* draw a true or at least well-justified causal graph representing the possible (i.e., not excluded) direct causal relations between the observed, the unobserved, and the imagined unobservable variables of interest. Moreover, based on evidence of selective influence, dissociation, or interference (Sternberg, 2001), we can sometimes legitimately infer the existence of distinct, *qualitatively* characterized latent structures, i.e., subsystems, modules, components, processes, or stages of the response process. Based on such patterns of results, we have reasons to believe, for example, that there are different types of memory or that there may be a bottleneck at the decision-making stage (Levy et al., 2006).

That is why the part of causal inference theory concerned with qualitative causal relations may be particularly useful in psychology. In Pearl's theory, qualitative causal relations are represented using *directed graphs*, i.e., graphs in which every connection, or *edge*, between two distinct *vertices* or *nodes* has a direction denoted by an arrowhead. For convenience, we also use *arcs*, i.e., two-way connections which indicate the possible existence of an unmodeled common cause. Due to the complications they are associated with, I will not consider here any *cyclic* graphs, i.e., graphs for which it is possible to return to the same node by going in the direction of the arrows.

At worst, causal graphs will make it easy to see that, given what is known about the sampling process, too many causal relations cannot be excluded, and some or all causal quantities of interest simply cannot be estimated, no matter the method. Such a result may be disappointing, but it is worth knowing because it discourages the formulation of unfounded conclusions and prevents giving undue weight to unsupported causal claims made by others. If it turns out that certain causal quantities can be estimated, it will only remain to determine the general form of the estimator (expressed in terms of unspecified distributions of observed variables) and find a good statistical approximation (usually a function of simplified or partial descriptions of the distributions of observed variables). In the simplest situations, this will only involve fitting some regression model.

The previously simulated process can be represented using the causal graph $X \rightarrow Y$. Unless stated otherwise, every arrow in a causal graph is interpreted as a *theoretical possibility* of the corresponding *direct* causal effect. By saying that a causal effect is direct, we do not mean that it is immediate: we only mean that it is not mediated by other modeled variables. Since, while maintaining the graph's meaning, one can label every arrow and arc as "unknown", marked edges do not need to be justified. Thus, every arrow or arc represents the lack of the corresponding causal assumption. It is the missing edges that correspond to the actual causal assumptions, and it is the missing edges that call for some justification because it is the missing edges that have categorical statistical consequences and sometimes imply that certain causal quantities can be estimated.

Unmodeled variables are usually not marked because one can infer the arrows that they emit. However, we must label unmodeled variables that may be statistically dependent. Except for spurious dependence resulting from conditioning on a collider, which I will discuss later, it follows from the meta-assumption that there is no correlation without causation that dependence between unmodeled variables is possible only when one such variable influences the other or when both have a common cause. Such dependence has important consequences, so we must mark it even when we do not mark other unmodeled variables. We do this using a bidirectional arc because dependence between two unmodeled variables implies that the corresponding modeled variables may have an unmodeled common cause. Thus, the graph $X \to Y$ is a simplified version of the graph $U_X \dashrightarrow X \to Y \dashleftarrow U_Y$, where the variables U_X and U_Y are assumed to be independent because no arc connects them (and no arc connects the variables X and Y in the simplified graph).

The relations represented by a causal graph are qualitative in the sense that nothing is assumed about the quantitative properties of the process just by accepting the graph. This allows the construction of a true or well-justified graph describing an arbitrary psychological study. In particular, unlike typical SEM[8] models (Blalock, 2018; Bollen, 1989; Duncan, 2014; Wright, 1921), which are linear and do not allow for interactive effects, in qualitative causal models we allow both linear and nonlinear relationships. Apart from the independencies implied by the missing arcs, we also do not assume anything about the distributions of unmodeled variables. Finally, when a variable is (potentially) directly caused by more than one variable, we allow for interactive effects.

The conclusions based on a causal graph will be valid *as long as no arrow or arc corresponding to an actual causal relation is missing*. In particular, when there are more arrows or arcs than are necessary – meaning that not all correspond to relations of real influence but no real causal relation is left unmarked on the graph – the conclusions will still be true, but perhaps the graph may have fewer testable properties, and it may allow for fewer causal inferences from data. The interpretation of arrows as theoretically possible direct causal relations also lets us include as modeled variables arbitrary theoretical constructs, i.e., unobservable variables that may or may not exist, without making the resulting graph false if such a variable does not exist[9].

Perhaps the most important part of the analysis of every causal graph is concerned with their *paths*, i.e., non-empty, finite sequences of adjacent arrows without repetitions, such that the arrows comprising the path do not have to go in only one direction[10]. The analysis of paths often comes down to making use of the properties of the *chain* $X \to Y \to Z$, the *fork* $X \leftarrow Y \to Z$, and the *collider* $X \to Y \leftarrow Z$. Remembering the properties of these three structures greatly facilitates the use of the graphical part of the theory. Since paths look like graphs but are, by definition, only parts of some (perhaps unspecified) graphs, from now on the reader will need to pay attention to whether the causal structures under consideration are paths or graphs.

If X and Z are connected by the chain $X \to Y \to Z$, then X and Z may – but do not have to – be statistically dependent. If X and Z are not connected by another collider-free path, then X and Z must be independent in every *stratum* of Y, i.e., in every subset of the population where only one specific value of Y naturally occurs

[8] I will refer to linear structural equation models in this way to distinguish them from more general structural models.

[9] For example, one can adopt the convention that variables representing nonexistent theoretical constructs are, in fact, arbitrary constants.

[10] The term "path" is not always defined as a sequence without repetitions and in general, a graph-theoretic definition may allow the sequence to be empty.

or is merely observed: for every y, if we look at the subset of random samples such that $Y = y$, we will see no systematic statistical dependence between X and Z. In other words, according to this chain viewed as a graph, for every y, X and Z are independent in the conditional distribution $p(X, Z|y)$; the same thing is expressed by the equation $p(Z|X, Y) = p(Z|Y)$ or, more concisely, by the expression $X \perp\!\!\!\perp Z|Y$. Finally, in each stratum of some descendant of Y in the graph of which this chain is a part, the observed statistical dependence between X and Z may be weaker. The causal effect will not change because stratification is, by definition, selective observation and, as such, does not alter the way the process *works*: it only changes *how we view the outcomes of the process*. To the extent that stratification on the descendant of Y causes the variability of Y to not be fully manifest in the data, the causal effect mediated by Y may also not be fully manifest, although it will remain unchanged.

All these properties of chains can be illustrated using simulations. For simplicity, the effects will be linear, all intercepts will equal 0, all slopes will equal 1, and each unmodeled variable will have a standard normal distribution. When interpreted as a causal graph, the chain $X \rightarrow Y \rightarrow Z$, for instance, can be instantiated as follows:

```
X = rnorm(n)
Y = rnorm(n) + X
Z = rnorm(n) + Y
```

To make the correspondence between the code and its graph easier to see, I removed the "bookkeeping" instructions (resetting the random seed and setting the number of simulated samples), I did not name the unmodeled variables, and I changed the order of the summed terms. Now we can see that when we statistically control for Y, the statistical effect of X on Z is not statistically significant:

```
confint(lm(Z ~ X + Y))
#              2.5%  97.5%
# (Intercept) -0.02  0.02
# X           -0.04  0.01
# Y            0.98  1.02
```

The `#` sign is interpreted as the beginning of a comment and causes R to ignore the text that appears on the same line after this sign. From now on, I will use comments to add the results of evaluations of instructions to the code and I will round the numbers to two decimal places. The `lm` function fits a linear model, and `confint`, by default, returns 95% confidence intervals for all regression coefficients. The 95% confidence interval for the slope of X includes 0, meaning that the slope estimate is not significant at the $\alpha = ,05$ level. The statistical effect of Y on Z is consistent with the causal effect of Y on Z, which is given by the slope of 1. Below, we can also see an example of something that should interest those psychologists who use mediation analysis, namely that statistically controlling a descendant of Y, which can be interpreted as a perfectly valid (although not “perfectly reliable”) measure of Y, works quite differently than statistically controlling Y itself:

```
V = rnorm(n) + Y
confint(lm(Z ~ X + V))
#               2.5%  97.5%
# (Intercept) -0.03  0.02
# X            0.46  0.52
# V            0.48  0.52
```

As can be seen, the confidence interval around the effect of V has nothing to do with the causal effect of Y. Moreover, the statistical effect of X remains significant, which – without knowing the structure of the process – could be incorrectly interpreted as a reason to reject the assumption of complete mediation.

Statistically, the fork $X \leftarrow Y \rightarrow Z$ behaves the same as the chain, i.e., the only testable consequence of both paths as graphs is $X \perp\!\!\!\perp Z|Y$, and conditioning on a descendant of Y may make the observed dependence between X and Z weaker. These two paths are then *observationally indistinguishable as graphs*.

The collider behaves in an almost opposite and, at the same time, counterintuitive manner, which is why the set of its properties is called Berkson’s paradox (Berkson, 1946). If X and Z are two independent causes of Y, then X and Z are, of course, independent, but they *may be dependent in some or all strata of Y*. For example, if we consider only the instances of two independent dice rolls, X and Z, for which the sum, Y, is even, where the sum can be thought of as being caused by X and Z, then if X is an even number, then Z must also be an even number. Thus, it will not be the case that $X \perp\!\!\!\perp Z|Y$. By looking through the strata of a variable at the distribution of its actual causes, we may see – and we usually will – a systematically distorted dependence between the causes. However, the only categorical testable consequence of the graph $X \rightarrow Y \leftarrow Z$ is $X \perp\!\!\!\perp Z$. At this point, the reader should be able to write code that generates samples from this kind of process to illustrate Berkson’s paradox. I encourage beginners to do this, as it is worth understanding the counterintuitive properties of this frequently occurring structure.

To use the part of the theory introduced so far, let’s consider the consequences of conditioning on a collider, called selection bias. Sampling in psychological studies is hardly ever random and usually involves collecting data from individuals to whom the researcher has convenient access. If we denote as Z the set of variables that interest the researcher, and we denote as X the set of variables such that, due to the sampling method, the samples tend to be specific in terms of the values of these variables, then *for every pair of variables V and W in Z, if V and W influence some variable in X, the observed statistical dependence between V and W will be systematically distorted because of the way the sampling process works*.

As an example, consider a study on the relationship between gender at birth (G) and general intelligence (I). Because, according to current knowledge, gender at birth is the result of an essentially random process, regressing almost any variable Y (including I) on G estimates the causal effect of G on Y, i.e., $p(Y|G = g) = p\big(Y|do(G = g)\big)$. However, suppose the sampling process is such that psychology students have a greater chance of taking part in the study than would be the case if the sampling were random. In that case, there is a serious problem that must be addressed: both intelligence and gender certainly have a (strong) causal effect on whether someone becomes a psychology student. Because of this property of the sampling process, the observed statistical dependence between gender and intelligence will be systematically distorted.

As paths without a collider can "naturally," i.e., without stratification, induce statistical dependence, they are called *active*; paths with a collider are called *inactive* because they don't have this property. Statistical dependencies resulting from Berkson's paradox are called *spurious*, while those resulting from forks are sometimes called *noncausal*, but I prefer to avoid this term.

A causal graph has the same meaning as the corresponding list of so-called *nonparametric*, in the sense of abstract or unspecified, *structural equations*. These equations are called structural because, unlike ordinary symmetric mathematical equations, they represent, as functional relations, asymmetric deterministic causal relations. For example, the graph $X \rightarrow Y \leftarrow Z$ expresses the same assumptions as the following nonparametric structural model:

$$X = f_X(U_X)$$
$$Z = f_Z(U_Z)$$
$$Y = f_Y(X,Z,U_Y)$$

where, as explained earlier, we can assume that $f_X(U_X) = U_X$ and $f_Z(U_Z) = U_Z$. Both the graph $X \rightarrow Y \leftarrow Z$ and the above structural model are two representations of causal assumptions: about the lack of a direct causal effect of X on Z, Z on X, Y on X, Y on Z, and the assumption that unmodeled variables are all independent.

Every structural model can be interpreted as an abstract specification of a computer program since the equal sign in such models denotes an assignment operation, which can be performed either by Nature in general or by that specific part of Nature that is a computer. The presence of the structural equation $Z = f_Z(X,Y,U_Z)$, for instance, means that the values of Z are created independently of any variable outside its parent (or argument) set $\{X,Y,U_Z\}$, and that Z may or may not depend on the variables that belong to this set[11]. About unmodeled variables in nonparametric structural models, we only assume that they have some joint distribution which is unspecified apart from the assumed independencies.

We analyze the properties of paths and graphs that describe the process resulting from an intervention in the same way as before the intervention, but we must first remove all arrows entering the variables that are intervened on. In this way, within one model (family) we can consider different processes or versions of the same process, which we cannot do in probability calculus. The definition of an intervention as an operation on the graph involving the removal of

[11] The definition of mathematical function allows a function to ignore its arguments.

arrows entering the variables subject to intervention corresponds to the structural definition, which requires replacing the right-hand sides of the relevant structural equations, i.e., the parts specifying how the values of the variables arise, with constants. The structural definition, in turn, corresponds to how I simulate the effects of interventions here.

An intervention is thus understood as cutting off the variable(s) subject to the intervention from natural or previous sources of variability. If any property of the process could change in a way other than implied by the – not necessarily known – causal model, the effect of the intervention would be undefined. That is why an intervention in Pearl's theory is, by definition, local, meaning that other structural functions are assumed not to change. Estimating a causal effect, therefore, involves estimating an effect of a theoretical and perhaps unattainable ideal of an intervention that is external and "surgical", i.e., perfectly selective.

An Example of Using Causal Inference to Analyze and Interpret Data from a Psychological Experiment

All that has been said so far can be applied to arbitrary research designs, forms of statistical dependence, and target causal quantities as long as these quantities are specified at the level of a population of individuals or of a population understood as hypothetical replications of the same kind of sampling process on the same individual.

Before discussing a typical sequence of steps in causal analysis, I must introduce the definition of an important relation which can be understood as a formalization of the notion of blocking the flow of information. We say that a set of variables S *d-separates* a *path* p between variables X and Y if p contains a collider such that neither the collider nor any of its descendants is in S (then p is inactive, and conditioning on S cannot induce a spurious dependence between X and Y due to this path), or p contains a chain or fork such that the middle variable is in S (then conditioning on this element prevents the flow of information through p, regardless of whether p is active). The d-separation criterion can thus be understood in terms of the previously discussed properties of chains, forks, and colliders. If X and Y are two non-empty and disjoint sets of variables, we say that S *d-separates sets* X and Y if it d-separates every path between any two variables such that one is in X and the other is in Y.

I will now briefly describe some typical steps in causal analysis using Sternberg's study on short-term memory search as an example (Sternberg, 1969). In this study, shortly after viewing randomly presented sets of stimuli they were supposed to memorize, participants classified target stimuli as new or old. The key randomized variables were the set size (SS) and whether the stimulus was old or new; the measures were reaction time (RT) and accuracy (ACC). For simplicity, I will only discuss the condition where the target stimuli were new.

In one of the models, called the serial model, Sternberg assumed that the unobservable response process involves searching memory by comparing the representation of the target, sequentially and in random order without repetitions, with the representations of stimuli stored in memory, and each comparison takes on average the same amount of time (μ_T). For the condition with new test stimuli, these assumptions can be expressed by a structural model containing the structural equation $RT = \sum_{i=1}^{SS} T_i + U_{RT} = SS\mu_T + \epsilon_T + U_{RT}$, where T_i is the time of the i-th comparison, ϵ_T is the sum of the deviations of comparison times from the mean of the distribution of comparison times μ_T, and U_{RT} is the total duration of all other stages of the process, such as encoding the target stimulus and generating the motor response. According to this model, the effect of set size on mean reaction time is linear, and the slope is equal to the mean comparison time because a new target stimulus requires all the elements stored in memory to be checked. Note that this model assumes the graph $SS \rightarrow RT \leftarrow T$, which is quite optimistic in that many arrows and arcs are missing.

The first step in a causal analysis may involve *dividing the modeled variables* into those that are *observed* (SS, RT i ACC) and *unobserved* (T).

The second step could be, for all pairs of modeled variables, *drawing all the arrows and arcs that cannot be excluded or deemed negligible.* To ensure that all pairs have been considered, we can list all the modeled variables according to temporal order, e.g., SS_1, T_2, RT_3, ACC_4; then, we can select the pairs (i,j) where $i = 1, ..., n$-1 and $j = i + 1, ..., n$, where n is the number of variables, i.e., (SS_1,T_2), (SS_1,RT_3) (SS_1,ACC_4), (T_2,RT_3), (T_2,ACC_4), (RT_3,ACC_4). Because this stage involves formulating theoretical arguments, the reader may disagree with what I am about to say and remove some of the edges I will mark on the graph. However, each such removal would require theoretical justification *from the reader* as by drawing arrows and arcs I will merely point out that something is not known.

The constructed graph will represent processes occurring during a single trial, so its application may require that the statistical analysis be conducted on non-aggregated data. If analyses were conducted on data averaged over trials, it would be necessary to consider some causal relations between trials that we could otherwise ignore. For example, correctness in trial t may influence the response process in trial $t + 1$ if participants can sometimes realize that the response was incorrect, which will be possible if, e.g., there is feedback. Mean reaction time and mean accuracy are deterministic functions of reaction times and accuracy in averaged trials, so the possibility of such a causal effect would mean that we must mark the arrow $ACC \rightarrow RT$, which, as we will see, we do not need to do if we analyze non-aggregated data.

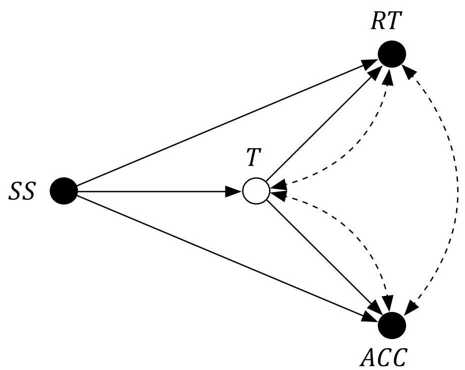
It is often convenient to start with the arcs. Because SS is randomized, no "real" arrow can enter it, but T should be connected by an arc with RT and ACC as it is hard to argue against the possibility that the efficiency or ease of memory search depends on other factors that may affect performance, such as, e.g., momentary distractions. Finally, for obvious reasons, such as motivation, strategy, learning, and fatigue, but also because of the aforementioned possible influence of the preceding response, RT and ACC should also be connected by an arc. SS may perhaps directly cause every other modeled variable because, e.g., participants may sometimes change how they perform the task depending on how difficult they perceive it to be. T can, of course, affect RT, and if there is external or internal time pressure, it can also affect ACC. Note that RT and ACC are two properties of the same reaction, with ACC also being a property of the stimulus class (new or old), but I did not mark this variable on the graph because I only consider trials with new target stimuli. In particular, RT is not the same as search time, and ACC is not the same as processing difficulty; RT and ACC are only observed consequences of the state of these and other variables, with RT being determined by the computer running the task, therefore RT and ACC cannot influence each other. Finally, due to the temporal order, these two variables cannot affect T. In this way, we obtain the graph shown in Figure 1.

The reader may disagree, but in my opinion the ontological status of the variable T is not obvious. For example, we do not know whether memory search involves comparing discrete representations, or if it is, perhaps, something akin to a more "fuzzy" process of accumulating evidence. However, we can adopt a way of thinking about some or all latent variables (here only T) that is similar to how we think about the edges; by thinking about some or all latent variables as mere theoretical possibilities, we can confidently assume that this graph is true.

The third step may involve *testing causal assumptions*. Testable consequences of a causal graph are precisely the (often conditional) independencies between observed variables implied by d-separation. It is important to bear in mind that for almost every causal graph, there will exist different graphs that have same modeled variables and are statistically indistinguishable. We say that pairs of such graphs are *statistically* or *observationally equivalent* or belong to the same *equivalence class*.

Figure 1

A graph representing theoretically possible causal relations between the set size SS, latent comparison time T, reaction time RT, and accuracy ACC in the condition with new stimuli in a short-term memory search task.



Statistically indistinguishable are, among others, all qualitative models that have the same modeled variables, that have the same *skeleton* – variables connected by an arrow or arc in one are also connected by an arrow or arc in the other – and that have the same *V-shaped structures*, i.e., structures where two direct causes (i.e., parents) of the same variable (node) are not connected by an edge (Verma & Pearl, 2022). By changing the direction of any arrow in the graph, we create a statistically indistinguishable model, i.e., a model that cannot be distinguished by observing all modeled variables (including the latent ones!) from the original model, as long as we do not remove an existing V-shaped structure or create a new one.

We can now rephrase the observation regarding the correlation between life satisfaction and yearly income using the precise and general language of causal inference theory. Instead of saying, "*The correlation between X and Y does not imply that X causes Y because this correlation may occur due to the existence of a common cause or due to the influence of Y on X,*" we can now say more generally, "*If X and Y are the modeled variables, then for every distribution of X and Y there exist processes that cannot be represented by the graph X → Y and that can generate this distribution.*" To show that inferring causation in a particular direction based solely on correlation is incorrect, it is sufficient, of course, to demonstrate one counterexample.

Returning to Sternberg's experiment, in this case no set of observed variables d-separates any pair of observed variables, so this graph has no testable properties meaning that every possible distribution of three variables can be generated by some process described by this graph. However, to be justified or valid this graph does not have to be testable; instead, it has to formally express all the theoretically possible causal relations that need to be considered when interpreting results.

Although it represents speculative causal assumptions, Sternberg's serial model is testable because there are possible patterns of results that are inconsistent with this model. However, every possible pattern of results consistent with this model can be explained by assuming completely different memory search processes. For instance, as Townsend et al. (1983) demonstrated through a different kind of formal analysis, as far as the possible results of this kind of study are concerned, the serial model is empirically indistinguishable from some theoretically acceptable parallel models. Even without such formal analyses, it can be noticed that – as is often the case in psychology – the pattern of results is relatively simple and, at best, moderately surprising, whereas the research goal is ambitious.

When discussing the issue of testing causal assumptions, it may be worth mentioning some less obvious consequences of randomization. In the case of stochastic processes, one cannot avoid the possibility of sampling error. It follows that expecting that truly[12] randomly assigned groups will be the same on all dimensions other than group membership is as unwise as expecting that a truly random sample will have all the features of the population (i.e., that the sample itself, not the sampling process, will be "representative"), or that two truly

[12] One of the reviewers noted that randomization might fail to be successful, i.e., that there might be a need to check if it worked as intended. However, in such cases, the causal structure of the potentially flawed randomization process should also be represented on the graph.

independently generated values will never be significantly correlated. Unless randomization is poorly done, the typically *untestable* assumption of independence of group membership status will be satisfied. As far as causally motivated statistical inference is concerned, this is all that matters because theoretical guarantees associated with statistical inference are properties of decision rules, i.e., they are *asymptotic*.

There are situations in which statistical control of variables observed before the randomized intervention may be justified and useful. However, if truly randomized groups differ significantly in terms of a variable, the values of which were determined by the data-generating process before randomization, it must be a result of sampling error. Introducing statistical corrections *for this reason* would be an example of relying on a *decision rule based only on sampling error*, i.e., it would be as helpful – asymptotically – as reading tea leaves. Statistical control of variables, the values of which were determined *after* the randomized intervention, may systematically distort the estimate of the total causal effect of the intervention, so it makes sense to do this only in special situations, such as in the context of mediation analysis when estimating the total causal effect is not the main goal.

The fourth and, in this case, the last step may be to determine *which target causal quantities are identified* and try to obtain good estimates of those that are. From the graph, it immediately follows that apart from the issues that may arise due to misspecification of a statistical model, regressing any observed variable other than SS on SS gives a correct estimate of the total causal effect of SS on that variable. At the same time, we see that "cleaning" the data by removing incorrect reactions may generate a spurious dependence between SS and RT. This may happen for two reasons: ACC is a collider of SS and T, as well as a collider of SS and every unmodeled cause of RT and ACC. This kind of data cleaning can then systematically distort how the causal effect of SS on RT manifests itself in the data. Moreover, a moment's thought is all one needs to see that we cannot safely assume that all correct reactions are generated by the process of interest because people are not robots performing simple tasks, and their reactions can be – and in an unknown proportion of trials certainly are –correct by mistake. Moreover, some or all correct responses may be generated by a process that is qualitatively the same as the process that generated some or all incorrect responses, but this common process may have different quantitative properties when it produces correct as opposed to incorrect responses, which may lead to systematically incorrect causal conclusions when analyzing only the correct reactions.

As should by now be obvious, without a good theory to begin with little can be learned about the complex and unobservable response process. It seems that well-justified conclusions from this study come down to the claim that, within certain limits, increasing the set size leads to longer reaction times in a roughly linear way and also to a higher probability of error. If we examine more than one person on more than one occasion, we will also certainly learn that these effects are inter- and intra-individually variable. For example, to a degree that depends on the person, given enough data we will certainly observe that performance tends to improve with practice.

A researcher willing to claim that more can be inferred from this kind of data would have to deal with the fact that, by using causal analysis, simulations can be created that illustrate alternative explanations of all the observed statistical effects. The only way to draw conclusions about the studied process that are stronger than those that a justified graph allows is to rely on a theory that implies sufficiently strong *quantitative* constraints. However, this requires having good reasons to claim that the theory is *approximately true as a description of the response process*, which is usually a priori known to be multidimensional, complex, unobservable, non-stationary, and idiosyncratic.

The above remarks may sound too strong, but they follow from two important theorems, one of which has been proven only recently. Firstly, we know that do-calculus is complete in the following sense: for any causal graph and disjoint sets of modeled variables X, Y and Z on this graph, where the set Z may be empty, the interventional distribution $p\big(Y|do(X = x), Z\big)$ is identified if and only if this fact can be established by using the three rules of do-calculus (see Shpitser & Pearl, 2008, where these authors also provide general identifiability conditions for counterfactual quantities). Sometimes, additional causal quantities may be identified by introducing quantitative assumptions, such as linearity, additivity, homogeneity, or monotonicity. However, when little is known about the quantitative properties of the studied processes, as is typically the case in psychology, relying on such assumptions will often be optimistic to the extent that borders on wishful thinking; strong causal conclusions will then be justified primarily due to these optimistic assumptions and to a lesser extent, if at all, due to the study design, the results, and to what is known to be likely true about the process. Secondly, we have recently learned that the counterfactual level is, in a technical sense, irreducible to the interventional level (Bareinboim et al., 2022), which, in turn, implies that certain questions about the quantitative properties of the data-generating process cannot be answered directly using randomization. This may perhaps mean that the quantitative part of a causal theory must be, to a certain degree, justified in a purely theoretical way, which, incidentally, does sometimes seem to happen in psychology (see, e.g., theories based on rational analysis in Chater & Oaksford, 1999).

Sternberg's study can be viewed in yet another way by drawing the graph presented in Figure 2. Because the validity of causal conclusions does not depend on the choice of modeled variables, for simplicity I have omitted response accuracy. I also have not marked any arcs because every path between any two observed variables, which here are only SS and RT, passing through an arc would have to contain a collider and would thus be inactive. As far as we only care about the alternative causal explanations of the observed distribution $p(RT|SS)$, we could then safely assume that this graph is true if it were not for the fact that, also for simplicity, I have optimistically excluded the arrow $TT \rightarrow M$, although it is certainly theoretically possible that a longer total memory search time could cause some memory information loss.

In psychological experiments, the goal is often to show that some latent psychological variable (here M) influences another latent variable (here TT), for example, that mood affects memory. That is why graphs resembling the one in Figure 2 will necessarily appear, even if only implicitly, in many such experiments.

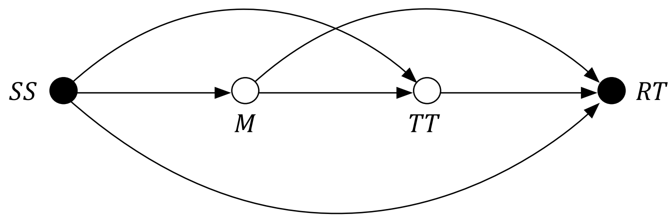
Sternberg assumed that SS could influence the number of items in memory (M), which is a latent variable, and that M could influence RT through the total search time (TT), which is also a latent variable. Apart from the consequences of the known properties of study design (randomization and temporal order), there is no reason yet to accept the optimistic graph $SS \rightarrow M \rightarrow TT \rightarrow RT$ and thus exclude any of the paths $SS \rightarrow TT \rightarrow RT$, $SS \rightarrow M \rightarrow RT$ and $SS \rightarrow RT$. So far, we have either exactly two categorical assumptions, i.e., the assumptions that M exists and that TT exists, or, if we want to entertain the very real possibility that not all theoretical constructs actually exist, all we have is an expression of an *intent* to study a particular possibly nonexistent causal effect of a variable that may not exist on another variable that may also not exist.

The only thing that randomization of SS guarantees is that the statistical effect of SS on any other variable Y is due to *some* directed path going from SS to Y. Like most measures used in psychology, RT can be under the *systematic* influence of many different factors, about which usually little is known (Borsboom, 2005; Millsap, 2012; Paulewicz & Blaut, 2022; Van Bork et al., 2022). Moreover, we can hardly ever safely assume that psychological interventions do not have any significant unintended consequences. Therefore, we often cannot exclude some or all additional paths marked on the graph in Figure 2, even in relatively simple experiments conducted under relatively controlled conditions.

As is often the case, the structure of the graph makes it easier to come up with alternative explanations. For instance, larger sets of items to memorize (SS) could make participants react more slowly, partly because upon seeing a larger set they become discouraged, or because during the presentation of successive stimuli the probability or degree of loss of focus tends to increase. Such effects could be mediated by search time ($SS \rightarrow TT \rightarrow RT$), by encoding time, or by something else ($SS \rightarrow M \rightarrow RT$, $SS \rightarrow RT$).

Figure 2

A simplified (see explanation in the text) graph representing theoretically possible causal relations between the set size SS, latent memory load M, latent total search time TT, and reaction time RT in a short-term memory search task.



The reader may have already wondered if it is worth trying to obtain evidence of modeled variables being statistically independent and remove the offending arrows or arcs based on such evidence. However, this requires accepting statistical assumptions that will not be jointly true, and, moreover, it may be far from clear whether they are jointly as close to being true as may be necessary to demonstrate statistical independence in a given context. It is usually at least likely that the presence of statistical independence implies the absence of active paths between the independent variables. In typical situations, however, demonstrating statistical independence involves showing that *a point hypothesis is true*, e.g., that the difference between means or the correlation is *exactly* 0; it does not suffice that a statistical test does not reject such a hypothesis. A useful alternative may be methods that rely on interval assumptions, but I will not write about them due to lack of space.

Confounding Paths and Ways of Dealing with Them

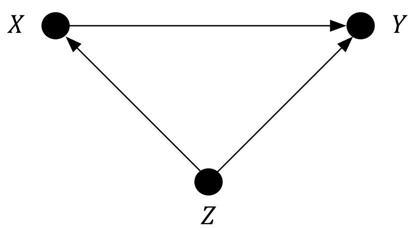
In psychology, the possibility of drawing justified causal conclusions tends to disappear when the cause of interest is not randomized. This also tends to happen when some variables are randomized, therefore the study is not simply observational, but it is observational with respect to the cause of interest. Although temporal order may justify the exclusion of some arrows, especially in psychology, it is often impossible to exclude the existence of unobserved common causes or to justify the assumption that their role is negligible.

Variables that are common causes and simultaneously serve as alternative explanations or co-occurring sources of statistical effects, which are interpreted as measures of the target causal effect, are known as *confounding variables*. However, it is not the confounding *variables* themselves that are problematic; instead, it is the presence of *active paths with a fork* that provide alternative explanations of observed dependencies or distort the way the target causal effect manifests in these dependencies. Moreover, neither of the two main ways of dealing with such paths, namely, back-door and front-door adjustments, requires the confounding variable(s) to be observed. Therefore, it is often better to talk about *confounding paths* rather than confounding variables.

Let's consider the graph in Figure 3

Figure 3

A graph with one confounding path ($X \leftarrow Z \rightarrow Y$) and one confounding variable (Z) with respect to $p(Y|do(X))$



… instantiated by the following process:

```
U_X = rnorm(n)
U_Y = rexp(n) - 1
U_Z = rbinom(n, size = 1, prob = 0.5)
Z = U_Z
X = U_X + 1 + 2 * Z
Y = U_Y + 3^Z + X^(Z + 1)
```

In R, the ^ symbol represents exponentiation. The `rbinom(n, size = 1, prob = 0.5)` instruction generates n pseudorandom samples with values 0 or 1 from the distribution $p(0) = p(1) = 0.5$. To illustrate the relative importance of statistical and causal assumptions in causal inference, this time the variable U_Y has a shifted exponential distribution (`rexp(n) - 1`); subtracting 1 ensures that this distribution has a mean of 0.

If we are interested in the effect $p(Y|do(X))$ rather than the joint effect $p(Y|do(X),do(Z))$, we need to deal with the confounding path $X \leftarrow Z \rightarrow Y$. If we wanted to estimate the joint causal effect of X and Z on Y, regressing Y on X and Z would be sufficient. Any regression of the form $p(Y|X,Z)$ estimates $p(Y|do(X),do(Z))$ here, since including Z as a predictor blocks the only problematic path without inducing any spurious dependencies between the observed variables[13].

Let's first estimate the mean of the distribution of Y in the situation $do(X = 0)$, but *only in the stratum* $Z = 0$. To achieve this, we can fit a false linear regression model to the subset of samples where $Z = 0$. This statistical model is false because the distribution $p(Y|X,Z)$ is a (sometimes shifted) exponential distribution (family). Still, in this stratum the confounding effect of Z cannot manifest because Z takes only one value. The statistical relationship between X and Y in the $Z = 0$ stratum may thus result only from the causal effect of X on Y in that stratum. So, even though the statistical model is false, the problem of confounding disappears. In other words, up to the approximation by a statistical model, regressing Y on X in the $Z = 0$ stratum says everything about the causal effect of X on Y in that stratum because it estimates the distribution $p(Y|X,Z = 0)$, and $p(Y|X = x,Z = 0) = p(Y|do(X = x),Z = 0)$. This is true for every x, but for simplicity I will focus only on the intervention $do(X = 0)$.

```
confint(lm(Y ~ X, subset = Z == 0))
#              2.5% 97.5%
# (Intercept)  0.96  1.04
# X            0.96  1.02
```

[13] The authors of the aforementioned handbooks on research methods and statistics for psychologists argue that a nonzero correlation between predictors, which is clearly present here, poses a significant problem when using linear regression, or maybe even that such a correlation is inconsistent with the assumptions of linear regression (it is not).

From the simulation code, it follows that when we observe $Z = 0$ and force $do(X = 0)$, R creates values of Y by calculating the value of the expression $3^Z + X^{Z+1} + U_Y$, which equals $3^0 + 0^{0+1} + U_Y = 1 + U_Y$, where U_Y has mean 0. The variable Y then has an exponential distribution with a mean of 1.

Since, in each stratum of Z, the statistical effect of X on Y is equal to the causal effect of X on Y, and the intercept in the fitted linear regression model represents the mean of Y when $X = 0$, the estimate of the intercept is an estimate of the stratum-specific version of the target causal quantity. This agrees with the fact that the true mean of the interventional distribution $p\big(Y|do(X = x),Z = 0\big)$, which is 1, is within the corresponding 95-percent confidence interval.

We can do the same thing with the stratum $Z = 1$. Since predictors in linear regression can be arbitrary functions of independent variables as long as the set of all predictors is not collinear (in the case of two variables, collinearity is the same as a correlation of 1 or -1), to adequately describe the systematic part of the statistical relationship in the $Z = 1$ stratum, we only need to create a variable equal to the square of X.

```
squareX = X^2
confint(lm(Y ~ squareX, subset = Z == 1))
#                2.5% 97.5%
# (Intercept)  2.91 3.02
# squareX      1.00 1.01
```

As is easy to verify, either by simulating the effects of the intervention or by calculating the exact theoretical value, the obtained interval estimate of the intercept contains the true mean of the interventional distribution $p\big(Y|do(X = 0),Z = 1\big)$, which is equal to 3. The expected value of Y in the situation $do(X = 0)$ is then sometimes 1 and sometimes 3, depending on which stratum of Z we are looking at. By multiplying 1 and 3 by the probabilities with which the two possible effects of $do(X = 0)$ occur, i.e., by $p(Z = 0)$ and $p(Z = 1)$ respectively, we obtain the expected value of Y when the intervention $do(X = 0)$ is forced on the *population*.

Generalizing this reasoning to the entire interventional distribution (not just the mean), arbitrary disjoint nonempty sets X and Y of modeled discrete variables, and an arbitrary intervention $do(X = x)$ yields the back-door adjustment. If there is more than one confounding path, it is necessary to block all of them simultaneously. A set of variables S such that *no descendant of (an element of the set) X is in S*, and S d-separates all the paths between X and Y that begin with an *incoming* arrow to (a variable in the set) X, is called a *sufficient set* (i.e., it is sufficient for estimating $p\big(Y|do(X)\big)$ using the back-door adjustment). The existence of such a set allows the back-door adjustment to be applied similarly to the way it was just done, with the role of Z being played by all variables belonging to S (by stratifying and summing or integrating over all of them):

$$\begin{aligned} p\big(Y|do(X = x)\big) &= \textstyle\sum_S p\big(Y|do(X = x),S = s\big)p(S = s) \\ &= \textstyle\sum_S p\big(Y|X = x,S = s\big)p(S = s) \end{aligned}$$

The *do* operator does not appear in the last expression. As we know, this is because, with respect to the causal effect on Y, in every stratum of S *observing* X is statistically equivalent to *intervening on* X. This is guaranteed by the d-separation of all confounding paths by the variables in S. Since the last expression contains only non-interventional quantities, the value of this expression can be estimated based on the results of an *observational* study: $p(Y|X,S)$ can be estimated using regression, and $p(S)$ can be estimated by fitting a parametric distribution. The obtained expression is a *universal generic estimator of the total causal effect of X on Y in any situation where a sufficient set exists, regardless of the quantitative properties of the process*. Note, however, that an integral will have to be used instead of a sum for every continuous variable in S.

This important generic estimator is called the back-door adjustment because the variables in a sufficient set block the "back entrances" to the variable(s) whose total causal effect we want to estimate. Variables controlled in this way do not have to be confounding variables, which is useful to know because not all confounding variables may be observed, or observing a blocking variable other than some confound may be less costly. Moreover, if a blocking variable is closer on the graph – in terms of the number of connecting arrows – to Y than the confounding variable, using such a variable to block the confounding path may improve the precision of the estimate, provided that using this blocking variable makes the percentage of variance "explained" in Y greater.

The frequentist approach to statistical inference may not be particularly helpful when we want to obtain *interval* estimates of this kind of quantity because the theoretical sampling distribution of the estimator obtained by applying the back-door adjustment will often not be known. In many such situations, however, researchers familiar with Bayesian inference may be able to obtain a good estimate by replacing $p(S)$ and $p(Y|X,S)$ with samples from appropriate posterior distributions. One must be careful, however, to approximate well the statistical effects of variables in the chosen sufficient set. In particular, successfully blocking confounding paths in psychological studies will often pose a severe challenge because nodes on the confounding paths will often represent unobserved variables, and statistically controlling an outcome of measuring a blocking variable has different consequences than statistically controlling the blocking variable itself.

If the chosen regression model does not capture the statistical effect of S on Y well, or if a *measurement* of a blocking variable is used instead of the blocking variable itself, residual dependencies, sometimes called residual confounding, may systematically distort the estimate of the causal effect of X on Y (i.e., they can lead to asymptotic bias). For example, if the process is a fork $X \leftarrow Z \rightarrow Y$, Z has the standard normal distribution, and Z_{01} is Z dichotomized according to the criterion $Z > 0$, then $X \perp\!\!\!\perp Y|Z$, but it is not true in general that $X \perp\!\!\!\perp Y|Z_{01}$. As is easy to illustrate through simulation, attempting to estimate the effect $p\big(Y|do(X)\big)$, which here is null, by applying the back-door adjustment using the regression $p(Y|X,Z_{01})$ and an estimate of the distribution $p(Z_{01})$ can lead to the erroneous conclusion that X has a direct causal effect on Y.

To check if the estimate derived earlier is correct, we can simulate the effects of the intervention and calculate the mean of Y, not for some subset of data where Z assumes only one particular value, but for the entire set of simulated values of Y, i.e., for samples from the interventional distribution of interest $p\big(Y|do(X=0)\big)$:

```
U_Y = rexp(n) - 1
U_Z = rbinom(n, size = 1, prob = 0.5)
Z = U_Z
X = 0
Y = U_Y + 3^Z + X^(Z + 1)
mean(Y)
# 2.00
```

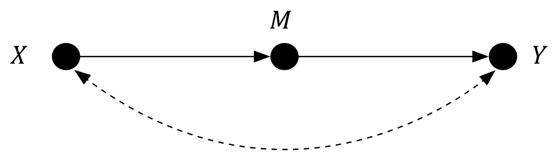
For comparison, naively, if the goal is to estimate the total causal effect of X, fitting linear regression of Y on X, or on X^2, or X and X^2, without accounting for the special causal role of Z, gives 95% confidence intervals around the intercept equal to [-1,61; -1,39], [0,79; 0,90], and [0,42; 0,57], respectively. In each case, the true interventional mean lies far outside the confidence intervals (if we measure the distance using the widths of the corresponding intervals).

The back-door adjustment or something closely related could be used, for example, in studies on the relative influence of genes (G) and parental traits or other properties of the family environment (F) on adult child traits (C). Due to temporal order, we can exclude the arrow $F \leftarrow C$, so this kind of study can certainly be described by the graph[14] $F \rightarrow C + F \leftarrow X \rightarrow C$, as long as we let X stand for all confounding variables, including genes. We cannot safely assume that $X = G$, but we can potentially obtain results indicating that this assumption is a good approximation. If, for example, based on the results of a large-sample study, we find that the statistical relationship between F and C becomes much weaker and close to nonexistent when we correctly control for the statistical effect of G, the conclusion that the statistical dependence between F and C is due in large part if not entirely to the influence of G will be justified.

If we have reason to believe in the assumptions expressed by the graph in Figure 4, we can take advantage of the fact that X blocks the only back-door path between M and Y.

Figure 4

Conditions enabling the estimation of the total causal effect of X on Y using the front-door adjustment



[14] The operation of adding graphs, which I introduced only for convenience, involves identifying vertices with the same names.

In every such situation, $p(M|do(X))$ can be estimated using regression, and $p(Y|do(M))$ can be estimated using the back-door adjustment, with $\{X\}$ serving as a sufficient set. By appropriately combining the two estimates, we obtain the *front-door adjustment*. Unfortunately, it is not easy to give examples of studies in psychology where this adjustment could be applied, and it is especially hard to find good examples of this kind in basic research. In basic research in psychology, the mediator will often be latent, mediation will often be partial, the measurement model of the mediator will probably be speculative and extremely simplified, and it will usually not be possible to rule out that the causal effects of X on M or M on Y are confounded (Rohrer et al., 2022). Therefore, readers interested in the front-door adjustment are referred to "The Book of Why" (Pearl & Mackenzie, 2021) or even to "Causal Inference in Statistics: A Primer" (Pearl et al., 2016) as, at this point, they should be sufficiently prepared to consult this excellent source.

Concluding Remarks

If correctly used, a theory of statistical inference can significantly lower the risk of drawing erroneous conclusions about *distributions*. A theory of causal inference plays a similar role at the *theoretical analysis and interpretation* stages, enabling the formalization of an essential part of substantive interpretation and facilitating the identification of all types of plausible causal explanations. Unfortunately, psychologists still often try to provide answers to causal questions, relying primarily, if not entirely, on statistical model comparison methods. I must admit that I, too, have made this fundamental mistake, sometimes more than once in the same publication (see, e.g., Paulewicz et al., 2007). This is essentially the same error of incorrectly accounting for the different status and role of causal and statistical assumptions, which I talked about at the very beginning, but beyond the familiar context of only two modeled variables, it is not as easy to detect.

The fewer arrows in the causal model, the simpler the corresponding statistical model will usually be because models with fewer arrows tend to have fewer free parameters, potentially making them more testable. Moreover, conclusions drawn by relying on causal models with fewer edges tend to be more interesting. This was apparent, for example, in the case of Sternberg's serial model. It would then seem that for these reasons, just as Occam's razor seems to dictate, in situations that raise doubts it may be better to remove arrows, arcs, or latent variables instead of leaving them. After all, simplicity and empirical testability, along with the generalizability of the associated statistical model, are qualities typically expected from theories or empirical hypotheses.

However, this approach serves to improve the *predictive*, not the *explanatory*, properties of a model. As the reader can now see by comparing using, e.g., the likelihood-ratio test, regressions $p(Y|X)$ and $p(Y|X,Z)$ fitted to the results of simulating statistically indistinguishable graphs $X \to Y + X \leftarrow Z \to Y$ and

$X \rightarrow Y + X \rightarrow Z \leftarrow Y$, whether one statistical model fits better than the other is logically independent of which of them provides interpretable estimates. That is why, for example, adding predictors because they seem to be in some way related to the variables of interest and choosing a regression model based on statistical tests in the hope of obtaining more interpretable estimates is a practice based on misunderstanding that will sooner or later, but inevitably, lead to entirely misleading conclusions (Cinelli et al., 2021).

The lack of a good theory can be sharply felt in psychometrics (see, e.g., Borsboom, 2005), which is the field concerned with how we can or should measure psychological latent variables. I may be able to write more about the causal-theoretic view of measurement of latent psychological variables in the next part of this introduction, but for now I would like to draw the reader's attention to one issue. Justifying the conclusion that a certain set of test items is under the systematic influence of only one latent variable by the fact that some factor-analytic model fits well or doesn't is not as productive an activity as it may perhaps seem to the many psychologists who appear to rely on this kind of reasoning routinely. The qualitative causal structure of every unidimensional factor-analytic model is not testable because the common factor in this generalized fork is, by definition, unobserved: *every* distribution of responses to test items can be generated by a process that has this structure, even if not every such distribution looks like it was generated by a *linear* process with this structure. The only testable part of these models is the assumptions about the linearity of effects and the normality of errors. However, the quantitative assumptions of the linearity of effects and normality of errors are introduced to simplify calculations or enable identification. In psychological testing, these assumptions are a priori known to be far from the truth simply because the responses are discrete and bounded, not to mention that the number of response categories is typically small. What is then being assessed in this way is only the *predictive performance of quantitative assumptions that are clearly false, all the while assuming an untestable and usually, at best, only weakly theoretically justified qualitative causal structure*. This pessimistic view already follows from the part of the theory presented so far, and the same can be learned from books on SEM models in which causality is taken seriously (e.g., Hoyle, 2012; Kline, 2015). That it is not even clear *where to look* for a satisfactory solution to the problem of systematic causes of error in the measurement of psychological latent variables can be learned from, among other sources, Millsap's excellent book on measurement invariance (2012), as well as from my and Blaut's modest contribution to this literature (Paulewicz & Blaut, 2022).

We know that do-calculus is complete, and no counterexamples showing that its axioms may be invalid have been found, at least to my knowledge. Moreover, the theory of causal inference is already developed enough that, for some important classes of problems, we know not only that this theory provides *some* solutions to these types of problems but also that it provides *all possible* solutions. There exist already perfectly usable and more or less fully developed, or emerging before our eyes, parts of the theory concerning mediation (Pearl, 2012), missing observations (replacing outdated classical methods, the requirements of which are so difficult to establish in practice that it is rarely known when they can be applied; see Mohan et al., 2013), integration of results of similar or only related observational or experimental studies (replacing as well as broadening the applicability of causally blind methods of meta-analysis; see Bareinboim & Pearl, 2016), ways of dealing with sample bias (Bareinboim et al., 2022), and other issues of critical importance to basic or applied research. The mathematical parts of research methodology in psychology, including the theory of how to plan and analyze research results and the theory of creating measurement instruments, assessing their measurement properties, and interpreting measurement results, were until recently based only on probability calculus and statistical inference theory. However, the most important methodological problems are causal, and their statistical aspect is only of secondary importance. And as the examples I have discussed here clearly show, relying on intuition in situations to which the theorems of causal inference theory apply is as reasonable as neglecting the theorems of probability calculus or the principles of logic.

References

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In R. Dechter, J. Halpern, & H. Geffner (Eds.), *Probabilistic and causal inference: The works of Judea Pearl* (pp. 507–556). ACM Books.

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352. https://doi.org/10.1073/pnas.1510507113

Bareinboim, E., Tian, J., & Pearl, J. (2022). Recovering from selection bias in causal and statistical inference. In R. Dechter, J. Halpern, & H. Geffner (Eds.), *Probabilistic and causal inference: The works of Judea Pearl* (pp. 433–450). ACM Books.

Bedyńska, S., Książek, M., & Cypriańska, M. (2012). *Statystyczny drogowskaz* [*A statistical signpost*]. Wydawnictwo Akademickie Sedno.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*(3), 47–53. https://doi.org/10.2307/3002000

Blalock, H. M. (2018). *Causal inferences in nonexperimental research*. UNC Press Books.

Bollen, K. A. (1989). *Structural equations with latent variables* (T. 210). John Wiley & Sons.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Brzeziński, J. (2022). *Metodologia badań psychologicznych* [*Methodology of psychology research*]. Wydawnictwo Naukowe PWN.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. https://doi.org/10.1016/s1364-6613(98)01273-x

Cinelli, C., Forney, A., & Pearl, J. (2021). A crash course in good and bad controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552

Duncan, O. D. (2014). *Introduction to structural equation models*. Elsevier.

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Galles, D., & Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science*, *3*(1), 151–182. https://doi.org/10.1023/A:1009602825894

Greenland, S. (2022). The causal foundations of applied probability and statistics. In R. Dechter, J. Halpern, & H. Geffner (Eds.), *Probabilistic and causal inference: The works of Judea Pearl* (pp. 605–624). ACM Books.

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford Publications.

Levy, J., Pashler, H., & Boer, E. (2006). Central interference in driving: Is there any stopping the psychological refractory period? *Psychological Science*, *17*(3), 228–235. https://doi.org/10.1111/j.1467-9280.2006.01690.x

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing System*, 26 (NIPS-2013) (pp. 1277–1285). Curran Associates, Inc.

Paulewicz, B., & Blaut, A. (2022). The general causal cumulative model of ordinal response. *PsyArXiv*. https://doi.org/10.31234/osf.io/e7a3x

Paulewicz, B., Chuderski, A., & Nęcka, E. (2007). Insight problem solving, fluid intelligence, and executive control: A structural equation modeling approach. In S. Vosniadou, D. Kayser, & A. Protopapas (Eds.), *Proceedings of the European Cognitive Science Conference 2007* (pp. 586–591). Psychology Press.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge University Press.

Pearl, J. (2012). The causal mediation formula – a guide to the assessment of pathways and mechanisms. *Prevention Science*, *13*(4), 426–436. https://doi.org/10.1007/s11121-011-0270-1

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2021). *Przyczyny i skutki. Rewolucyjna nauka wnioskowania przyczynowego*. Copernicus Center Press.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna. https://www.R-project.org/

Rohrer, J. M., Hu¨nermund, P., Arslan, R. C., & Elson, M. (2022). That's a lot to PROCESS! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095827

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *10* (469), 322–331. https://doi.org/10.1198/016214504000001880

Saville, D. J., & Wood, G. R. (2012). *Statistical methods: The geometric approach*. Springer Science & Business Media.

Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, *9*, 1941–1979. https://doi.org/10.5555/1390681.1442797

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457. https://www.jstor.org/stable/27828738

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*(1), 147–246. https://doi.org/10.1016/s0001-6918(00)00045-7

Townsend, J. T., &Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. Cambridge University Press.

Van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. https://doi.org/10.1037/met0000521

Verma, T. S., & Pearl, J. (2022). Equivalence and synthesis of causal models. In R. Dechter, J. Halpern, & H. Geffner (red.), *Probabilistic and causal inference: The works of Judea Pearl* (pp. 221–236). ACM Books.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, *20*, 557–585.