Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 15 (2). 2025. 227-249. Published online: 30.06.2025
https://doi.org/10.14746/ssllt.48231
http://pressto.amu.edu.pl/index.php/ssllt

Causality and ability beliefs: An introduction to confounders and colliders

Ali H. Al-Hoorie ✉
Royal Commission for Jubail and Yanbu, Saudi Arabia
https://orcid.org/0000-0003-3810-5978
hoorie_a@rcjy.edu.sa

Phil Hiver
Florida State University, Tallahassee, USA
https://orcid.org/0000-0002-2004-7960
phiver@fsu.edu

Abstract
Causal inference is a fundamental goal of many research endeavors, including scholarship in the field of language education and learning. Randomized controlled trials are considered an ideal design to test causal claims, but not all claims can be subjected to experimental treatment due to ethical and practical constraints. In this article, we provide an overview of the conditions under which causal inference may be made from observational data. This includes recognition of the role of confounders and colliders; the former are common causes of the independent and dependent variables and must be controlled, while the latter are common effects and must not be controlled. We illustrate these ideas with two examples involving ability beliefs and demonstrate them through directed acyclic graphs. We discuss the implications of this approach to causal inference from observational data, specifically in individual differences in language learning research, highlighting the need for explicit modeling of causal relationships and the risk of the atheoretical inclusion of variables, whether as controls, predictors, or covariates.

Keywords: DAG; d-separation; substantiation; overcontrol bias; endogenous selection bias

Any claim coming from an observational study
is most likely to be wrong
(Young & Karr, 2011, p. 116)

1. Introduction

Statistics textbooks typically tell readers that there are two options when dealing with data: descriptive statistics and inferential statistics. Descriptive statistics provide a concise summary of the data at hand (e.g., with means and standard deviations), while inferential statistics aim to come up with conclusions about the population from which the sample was selected (e.g., using *t*-tests and ANOVAs). While the descriptive-inferential distinction is useful, it does not clearly reflect the different goals of research from many paradigms and may therefore mislead unwitting researchers. Generally speaking, the goals of research include description, prediction, and causal inference (e.g., Hernán et al., 2019). While description is essentially the same in the two classifications, inferential statistics are further divided into those that are limited to prediction and those that extend to causation. At the prediction level, a researcher might find that a certain emotion (e.g., anxiety, happiness) is associated, either positively or negatively, with performance in a language task. At the causal level, however, the researcher would be advocating a counterfactual that if happiness or anxiety were absent (or if we intervened and changed their values), performance would have been different. This is an important distinction.

Although both prediction and causal inference are useful for different purposes, causal inference involves additional assumptions and is therefore harder to come by. These assumptions have to do with the data generating process, and therefore must come from outside the data. In other words, both description and prediction represent a data *reduction* process, while causal inference requires combining the data with knowledge about the world. This is why experimental designs permit causal inference since the researcher intervenes and influences how the data are generated. Thus, two datasets with the same set of variables can lead to different conclusions if one comes from an observational study and the other from an experiment. The observed association between a certain emotion and task performance may indeed be causal, or it might turn out to be an artifact of a third variable (e.g., self-efficacy) that is exerting a causal influence on both.

Clearly, making sound causal inferences is a crucial element of all quantitative individual differences (IDs) research in second language (L2) learning – particularly the case in language motivation research, where we locate much of our own work. Indeed, if a researcher reports that they obtained a correlation of, say, .40 between two variables, many readers will not think much of this finding if they consider this correlation simply spurious or that the actual (causal) relationship, if any, is much smaller. This is why statements like "The ideal L2 self was consistently found to *correlate* highly with the criterion measure (Intended effort), explaining 42% of the variance, which is an exceptionally high figure in motivation studies" (Dörnyei, 2009, p. 31, emphasis added) leave many readers unsatisfied, as they do not clarify the extent to which this correlation underlies a genuine causal process or merely a spurious association. Indeed, an unusually high correlation may invite an equally high risk of a spurious association. Additionally, assuming a causal relationship, readers would also wonder about the direction of causality, especially with more pronounced preliminary evidence suggesting a reverse causal process with effort potentially increasing the ideal L2 self over time (see Hiver & Al-Hoorie, 2020a).

While the correlation-is-not-causation mantra is well-established in the field in theory, practice suggests otherwise. Researchers studying IDs in language learning routinely present correlation and regression tables listing numerous variables with stars representing significance levels. This stargazing syndrome (McElreath, 2020), coupled with a scarcity of intervention research (Al-Hoorie et al., 2022; Lamb, 2017) and the customary practice of devoting entire manuscript sections to pedagogical implications from these observational results (Al-Hoorie et al., 2021), suggests that the field only pays lip service to the concept that correlation is not causation. Attempting to inform practice implies that the researcher, at least implicitly, presupposes a causal relationship underlying their findings. In reality, correlational results, even if consistently replicable with the same designs, usually do not hold up when examined experimentally. This is because there are usually many more correlational than causal relationships in any system (e.g., voodoo correlations, Fiedler, 2011; crud factor, Orben & Lakens, 2020; piranha principle, Tosh et al., 2021). For example, in an ambitious metascientific undertaking, Young and Karr (2011) reviewed 52 claims from observational research and found that not a single one of them withstood randomized controlled testing, and five even showed the opposite pattern.

More careful researchers avoid the C-word for other euphemisms like association, link, relationship, and variance explained (Grosz et al., 2020; Hernán, 2018). This practice can be misleading. In many cases, researchers and their readers are actually interested in the causal relationship underlying their findings, not just the extent to which some variables are associated. Avoiding being explicit about the causal goal of research (e.g., because the design is observational) conflates the means and ends of research. While it is justifiable to use associational language in the results section of an observational study, it becomes counterproductive in the introduction and discussion sections because it introduces unnecessary ambiguity as to the purpose of the study, and readers are in any case likely to jump to causal conclusions even if these are not warranted. Being transparent about one's causal goals, in contrast, forces researchers to be explicit about their assumptions, to recognize the limitations of their designs, and for the community to have an informed discussion about the potential implications of the findings.

At the same time, we have to admit that interventions that permit clear cause-effect inferences involve numerous practical and ethical challenges, which have made them "a tiny proportion" (Lamb, 2017, p. 334) of the total research into IDs in language learning. While conducting more experimental research is laudable, this is not always feasible (Hiver & Al-Hoorie, 2020b). The purpose of the present article is therefore not to call for more experimental designs, but to discuss the assumptions behind making principled causal inferences from observational research. Indeed, if the purpose of most research is to uncover genuine causal relationships (not mere associations with suspect causality), and if the only way to uncover these causal relationships is through randomized controlled experiments, then a lot of research will lose its value. There is, however, no reason for this to be the case. The science of causality has made great advances in the last few decades (Pearl, 2009; Pearl et al., 2016), attempting to resolve centuries-old debates and ambiguities surrounding causality and causal inference. This emerging body of research helps scholars think more clearly about data coming from observational designs and the assumptions needed to deduce causality from these data, which requires making those assumptions explicit and transparent (Rohrer, 2018). In this article, we introduce the role of *confounders* and *colliders* in making causal inferences. Throughout, we illustrate concepts with (hypothetical) examples to make the ideas less abstract. A recurring theme in this article is that the researcher is expected to develop an explicit and transparent model of the causal links between their variables while designing the study, one that allows other researchers to critique and build on it, rather than mechanistically including more and more variables in empirical research (Wysocki et al., 2022).

2. Confounders: The case of too few controls

A primary obstacle that prevents inferring causality from observational data is confounding. A confounder is basically a common cause that introduces a previously nonexistent association between two variables, or spuriously increases the magnitude of an existing one. More technically, a confounder is a variable that makes the observational distribution different from the interventional distribution. In order to better understand the intuition behind a confounder, imagine that a researcher computes the correlations among a collection of variables and is then surprised to find a positive correlation between shoe size and reading ability. Intrigued, the researcher collects new data and finds that this relationship replicates reliably. Of course, the more plausible interpretation is not that larger shoes lead to better reading competence, but that the sample comes from learners of different ages. Clearly, this association is spurious and not causal. However, a researcher asking about whether the independent variable (e.g., shoe size) has anything to do with the dependent variable (e.g., reading ability), without removing the effect of the confounder (age in this example), will find that shoe size and reading ability indeed appear to be empirically associated (see Figure 1, panel a). This is because as children age, their shoe sizes increase, and their reading ability also tends to improve. As shown by the dashed curve in Figure 1, panel a, the (non-causal) association is transmitted from the independent variable through the confounder and reaches the dependent variable (note that this non-causal transmission can travel against the direction of the arrows, which may sound confusing at first). This is called a *back-door path*.

Experimental and observational designs handle this third-variable problem differently. In an experiment, randomization is intended to minimize the risk of this problem in that the independent variable is not systematically associated with that third variable. In our example, the researcher would randomly give children shoes with different sizes regardless of their age, so that age is no longer related to shoe size. This is illustrated by the missing arrow from age to shoe size in panel b of Figure 1 (technically, what connects two variables is called an “edge,” while a “path” is the route where causality flows, possibly passing through multiple variables). This is why removing an arrow is a stronger assumption than including it (because, again, correlations are everywhere). In an observational study, however, the role of the researcher is to first identify that confounder and then remove its effect in order to exclude the non-causal association resulting from it. Removing the confounder may be accomplished by analysis through statistically controlling for it (also called conditioning on it and adjusting for it) or by design through stratifying the sample according to that confounder (in this article, we use “control” as an inclusive term). In this way, the researcher “blocks” the backdoor path and is able to estimate the path from shoe size to reading ability, as shown in panel b of Figure 1.

In many cases, controlling for confounders is the only option researchers have, as many variables are not amenable to experimental manipulation. For example, a researcher may be interested in examining whether an older sibling’s language aptitude has a causal impact on their younger sibling’s aptitude level (e.g., through socializing). Computing the correlation between the levels of aptitude in a sample of sibling pairs will not yield a precise estimate of the intended causal relationship because, simply, this correlation could be at least partly due to shared genetic variance. In order to avoid this problem, the researcher would need to remove the effect of genes. Of course, removing the effect of genes might be too difficult, so the researcher may decide to use parents' language aptitude as a proxy, imperfect as it is (see Figure 2). Controlling for a descendant of a confounder partially controls for that confounder. Of course, other researchers might think of other potential confounders of this relationship. These additional confounders can be tested in an improved model, and this is how research accumulates.

In order to further appreciate the risk of confounders, consider panel a of Figure 3. In this example, supportive teaching is hypothesized to improve student learning, and this process is further hypothesized to be mediated by students studying more hours. The researcher may decide to control for study hours in order to, first, test the hypothesis that study hours are a mediator and, second, estimate the direct effect of supportive teaching that is not mediated by study hours, if any. However, by controlling for study hours, a confounder is now introduced (MacKinnon & Pirlott, 2015). As shown in panel b of Figure 3, if a variable like conscientiousness has a causal effect on both study hours and achievement (Meyer et al., 2022), a back-door path will open, thus biasing the estimate. To block this back-door path, conscientiousness and all other confounders must also be controlled. Notice that this problem occurs even if the study is an experiment, as manipulating supportive teaching still does not affect path b in panel b of Figure 3 (study hours are also a collider; see later). This demonstrates the need to formulate clear conceptual and analytical models in an a priori manner and carefully consider potential confounders before causal inferences can be defensible.

In contrast to the back-door approach, researchers can also utilize the *front-door approach* (Pearl, 1995). The front-door approach permits causal inference even if it is not possible to control for confounders. Here, the researcher needs to identify the variable that mediates the relationship between the independent and dependent variables. This mediator must satisfy the front-door criterion, which requires that it fully mediates that relationship, and that there are no (uncontrolled) confounds either between the independent variable and the mediator or between the mediator and the dependent variable. This approach therefore blocks all back-door paths caused by the unmeasured confound. As an example, consider the effect of class size on student achievement. There are confounders (e.g., socioeconomic status, educational policies, etc.) that create a back-door path, consequently making the correlation between class size and student achievement non-causal. Typically, therefore, the researcher would need to conduct an experiment manipulating class size. With the front-door approach, however, the causal effect may be calculated from an observational study. The researcher needs to identify the mediator between this relationship, possibly teaching quality (see Figure 4). The researcher calculates the correlation between class size and teaching quality and between teaching quality and student achievement. The estimate of the total causal effect can then be obtained by combining these two correlations. Research by Glynn and Kashin (2018) has demonstrated that this approach provides estimates that closely approximate experimental results. The missing arrows in Figure 4 also explain the required assumptions – that the unknown/unmeasured confounders do not affect the teaching quality, and that no confounders exist between the independent variable and the mediator or between the mediator and the dependent variable – or that these confounders are controlled for. Thus, this approach forces the researcher to make their model and its assumptions explicit and transparent for the field to critique and build on.

A third method to make causal inferences without (direct) experimental manipulation is through an instrumental variables approach (Angrist et al., 1996; Hiver & Al-Hoorie, 2020b). This approach requires the identification of a variable that affects the independent variable of interest, but one that is not associated with the dependent variable or any confounder affecting the independent and dependent variables. For example, the researcher might be interested in finding out the causal effect of teaching experience on student achievement (see Figure 5). The mere correlation between teaching experience and student achievement is non-causal due to confounding, and the researcher may not be able to experimentally manipulate teaching experience. With an instrumental variables approach, the researcher identifies an instrument, such as available teaching positions in a certain geographical area. Availability of teaching positions is typically an administrative decision governed by financial or policy considerations. The availability of positions in different locations requiring different levels of teaching experience is therefore not expected to be related to student achievement or a confounder between teaching experience and student achievement. Based on these assumptions, the association between teaching experience and student achievement may be interpreted as causal.

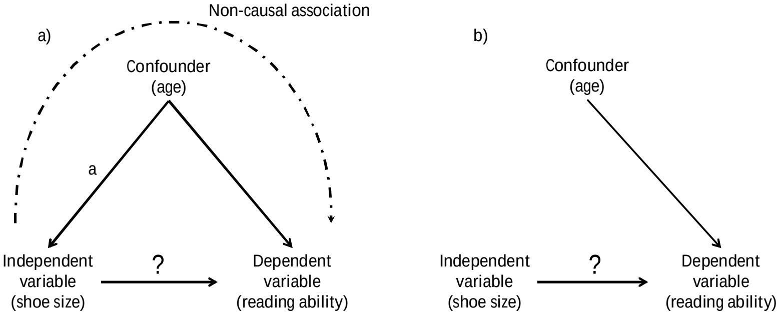

Figure 1 a) back-door path due to confounding; b) removing the effect of the confounder through experimental manipulation

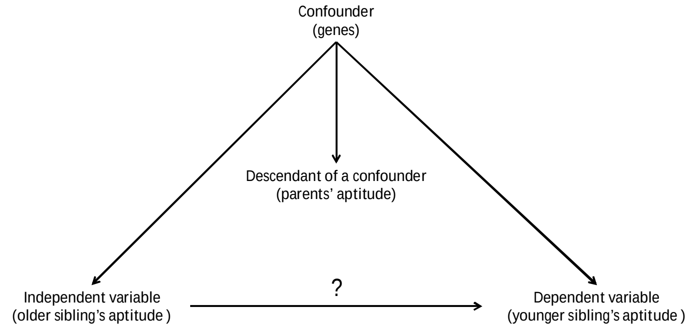
Confounder
(genes)
Descendant of a confounder
(parents' aptitude)
Independent variable
(older sibling's aptitude)
?
Dependent variable
(younger sibling's aptitude)

Figure 2 Controlling for a descendant of a confounder partially controls for that confounder

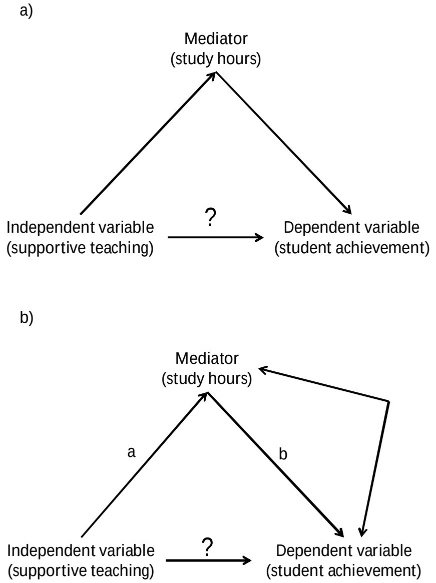

Figure 3 a) mediation model; b) confounder introduced because of the inclusion of a mediator

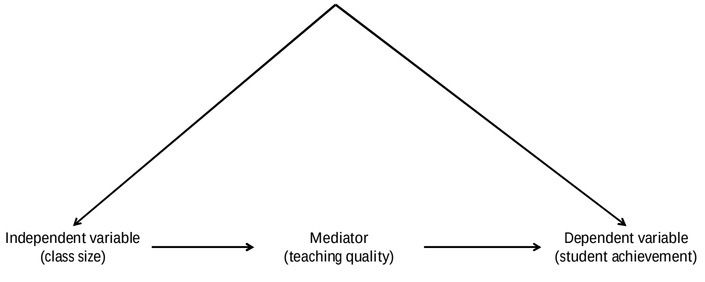

Figure 4 Teaching quality opening a front-door path if relevant assumptions hold

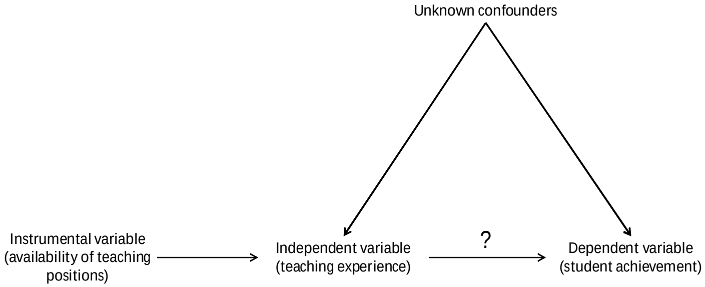
Unknown confounders
Instrumental variable (availability of teaching positions)
Independent variable (teaching experience)
?
Dependent variable (student achievement)

Figure 5 Availability of teaching positions functions as an instrumental variable assuming relevant assumptions hold

3. Colliders: The case of too many controls

If confounders prevent unbiased causal inference and should therefore be controlled, this might suggest that it is wise to try and control for everything the researcher could get their hands on – just in case. This more-the-merrier conception, where one believes that adding more controls to the model will improve inference, has been described as a methodological urban legend (Spector & Brannick, 2011) and as a practice that creates causal salad (McElreath, 2020). Indeed, many analysts feel frustrated by the common experience that adding just one more variable can make a huge difference in the results, sometimes even changing the sign (from negative to positive or vice versa) altogether. This issue is complicated further by the unhappy fact that causally incorrect models can make *better* predictions (McElreath, 2020). Thus, a model that can, consistently and replicably, explain more of the variance might still be a wrong model. Therefore, on the one hand, adding a variable (even if apparently irrelevant, such as shoe size [see Figure 2]) can lead to confounding through opening up a back-door path, thus introducing a spurious, non-causal association. On the other hand, adding a variable with the intention to control for it can open up a previously closed back-door path, again introducing a non-causal association. In this section, we discuss this problem of bad controls (Cinelli et al., 2022). We start with the idea of colliders before moving to mediators and posttreatment variables.

While confounders introduce omitted variable bias, colliders lead to included variable bias. It might, at first, seem counterintuitive that adding another variable (as predictor, control, or covariate) with the intention of improving the model can instead end up damaging it. To better understand the intuition behind this, consider an extension of the hypothetical aptitude example above. Imagine this time that a researcher is interested in finding out whether a husband's language aptitude has a causal effect on the wife's aptitude level. Unlike the siblings example, couples may not have shared genetic variance that might work as a confounder, so the researcher looks for other potential sources of confounding (e.g., a certain dimension of personality that attracts couples in the first place). Then the researcher might think, as there are no shared parents to control here, why not control for children's aptitude instead? This is a very bad idea. One way to look at this situation is that removing variance in a descendant of the dependent variable removes some of the very variance the researcher is targeting. Another way to look at it is that controlling for this variable opens up a previously closed back-door path, thus allowing non-causal association to flow through it. Graphically, while a confounder is represented with a fork (e.g., Figure 1, panel a), a collider is represented with an inverted fork (see Figure 6). A collider, therefore, is a common effect. Non-causal association (dashed curve in Figure 6) does not flow through a variable with an inverted fork unless it is controlled (i.e., it is blocked by default). In short, while researchers should control for parents (confounders), they should not control for children (colliders).

To further demonstrate the complexities involved, imagine that a researcher is interested in testing the effect of some form of supportive teaching on student achievement. The researcher examines the relationship between these two variables (it does not matter whether the study is experimental or observational for the purpose of this example) but is disappointed to find no association. The researcher might hypothesize that students' "happiness" might somehow be related to this relationship, and so they decide to control for this variable, and then they indeed find a relationship. What could possibly go wrong? Controlling for happiness opens up a back-door path if it is a collider on the path between supportive teaching and the unmeasured variable well-being. If this happens, it will in turn introduce the unmeasured confounder income, which has a causal influence on both well-being and student achievement (Killingsworth, 2021; see Figure 7). These additional variables are outside of the researcher's radar altogether and were therefore not included in the study design. This situation underscores the need to carefully consider the ramifications of selecting a variable to control. Sometimes controlling for a variable to avoid one source of bias can lead to another source of bias, in a process called *butterfly bias* (Ding & Miratrix, 2015).

A topic related to collider bias is controlling for mediators. A mediator is an intervening variable that explains the mechanism of the effect of the independent variable on the dependent variable. Controlling for a mediator may be helpful for the purpose of calculating direct and indirect effects, though (as shown in Figure 3) this requires strong assumptions, such as the absence of confounding (MacKinnon & Pirlott, 2015). Problems arise when the researcher is not aware that a variable acts as a mediator and controls for it. The unwitting researcher would be controlling for the very effect they are after. Due to overcontrol bias, the effect will completely disappear in the case of full mediation or will be reduced in the case of partial mediation. In hindsight, it might seem common sense that controlling for a mediator is problematic, but some situations can be tricky. Consider this example: A researcher is interested in the effect of supportive teaching on student achievement. The researcher knows that this effect is mediated by student motivation, so they are careful not to control for this mediator. But the researcher controls for class participation. Hypothesizing that the effect of supportive teaching might be different for those who actively participate in class and those who do not, the researcher controls for class participation either through entering this variable as a predictor in the regression model or by splitting the sample and comparing high-participators and low-participators. This is a bad idea. As shown in Figure 8, because it is a manifestation of behavioral engagement (Hiver et al., 2021; Zhou et al., 2021), class participation is a descendant of the mediator in this relationship. Controlling for a descendant of a mediator constitutes partially controlling for the mediator itself, again removing the very effect the researcher is investigating. This again demonstrates the need to explicitly state the reasons behind adding controls in the model or dividing the sample into subgroups.

A final and closely related point we discuss in this section is the idea of conditioning on (i.e., controlling for) posttreatment variables. In the above example about controlling for class participation, it matters *when* class participation data were collected: before or after they were influenced by the independent variable supportive teaching. At first, it might not be clear why the timing of participation data collection is relevant, but consider the following explanation. In the sample, there are teachers who provide supportive teaching and those who do not. Within each case, as a result, there will be students who exhibit high participation and those who exhibit low participation. If the researcher splits the sample (or computes interaction effects) based on high versus low participation, the groups will *not* be comparable. Low participation after supportive teaching is not the same as low participation after non-supportive teaching. The former points to students with low motivation who refrained from participating *despite* supportive teaching, while the latter possibly includes students with *higher* motivation who did not manifest in participation due to lack of supportive teaching. Comparing these groups is similar to comparing apples and oranges. This problem of conditioning on posttreatment variables applies equally to experimental (Montgomery et al., 2018) and observational (Acharya et al., 2016) research, perhaps more so in the latter case. Avoiding this problem in the case of observational research requires collecting participation data before these data are influenced by the independent variable supportive teaching (e.g., during the previous semester). A similar problem occurs with participant exclusion, noncompliance, and attrition happening during the course of the study, which leads to endogenous selection bias (Elwert & Winship, 2014). As controlling for posttreatment variables is useful in only limited cases (e.g., Cinelli et al., 2022), researchers need to carefully plan not only what variables and controls they need but also when these should be measured.

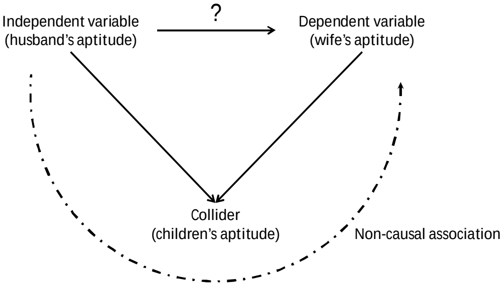

Figure 6 Controlling for a collider opens up a back-door path

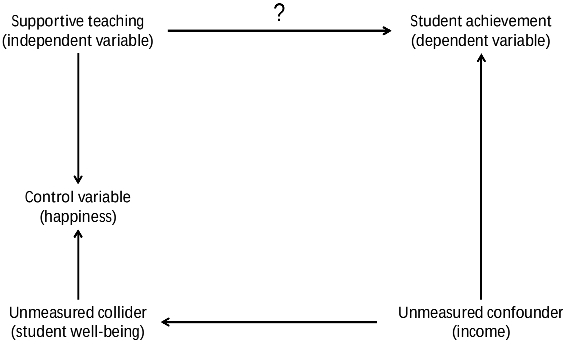

Figure 7 Controlling for one variable (happiness) could open up a back-door path

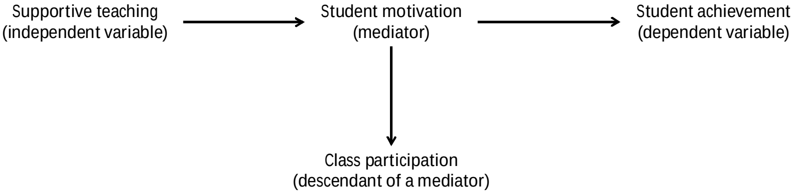

Figure 8 Controlling for a descendant of a mediator constitutes partially controlling for that mediator

4. Discussion

This article has reviewed an approach to making causal inferences from observational data in a principled fashion. In most situations, authors and their readers are, explicitly or implicitly, interested in the causal implications of a set of findings, not just mere associations that might end up being spurious. However, because researchers are usually aware of the limitations of their designs, causality has become "a dirty word that respectable investigators do not say in public or put in print" (Hernán, 2018, p. 616). This mindset muddies the waters, goes against transparency values, and prevents other researchers from zooming in on these causal claims, subjecting them to careful testing, and elaborating and expanding on them. We therefore urge fellow researchers to address causal claims explicitly rather than shying away from this issue. The approach presented here falls under the rubric of *d*-separation, where *d* stands for directional (Pearl, 2009; Pearl et al., 2016; for an accessible treatment, see Hayduk et al., 2003). Causal inference may be made to the extent that two variables are *d*-separated, which requires that all confounders are controlled to close back-door paths and that no colliders are controlled to prevent opening up back-door paths that are already closed. Thus, while it is true, as stated in the epigraph to this article, that most claims coming from observational studies are false in the sense that they do not underlie the causal relationships researchers are typically interested in (as correlations are everywhere), *d*-separation with an explicit model offers a formal approach to testing causal claims. Randomized controlled experiments are not the only approach in the researcher's arsenal, and hence researchers should think beyond experiments (Diener et al., 2022).

The graphical representations we used in this article are known as directed acyclic graphs (DAGs). DAGs offer a simple but effective way to represent causal models (e.g., Tennant et al., 2020). DAGs are "acyclic" in that a variable cannot cause itself. Two variables may indeed have a mutual causal relationship, but the temporal dimension will make each a different variable ("anxiety at time 1" affects performance, which in turn affects "anxiety at time 2"). Furthermore, although the arrows linking variables are straight, DAGs are nonparametric. A DAG is simply a qualitative representation of the direction of causality, and thus it does not make assumptions regarding the sign of the relationship (positive or negative), its magnitude (small or large), its certainty (deterministic or probabilistic), its structure (simple or complex), or its shape (linear or nonlinear). It is a common misconception that a straight line means linear (e.g., Dörnyei & Ryan, 2015, p. 202). Whether the nonlinear relationship is exponential, logarithmic, polynomial, or follows any other pattern, and whether the interaction is nonlinear in that the relationship between two variables is not constant but depends on the level of a third variable, all this can be modeled with a DAG. We encourage scholars in the field to begin formally modeling causal claims using DAGs. For convenience, DAG generation can also be automated using open-source software such as DAGitty (www.dagitty.net; Textor et al., 2016) and Causal Fusion (www.causalfusion.net; Bareinboim & Pearl, 2016).

In this article, we also argued that adding one more variable to the model can sometimes have a damaging effect. Sometimes adding a variable produces a confounder bias (observed association becomes non-causal because the added variable introduces a previously nonexistent back-door path; see Figure 3, panel b), at other times, it leads to a collider bias (observed association also becomes non-causal because the added variable opens a previously existing but closed back-door path) and in some other times it results in overcontrol bias (removing the very effect the researcher is targeting). An unmeasured/unobserved variable can also damage the precision of the estimation of the causal effect as long as it is not included in the model and controlled, even if the researcher is unaware of that variable or its role (Tennant et al., 2020). These dynamics are relational, in that the very same variable can have different effects depending on what other variables are included in the model. Drawing a DAG forces the researcher to be explicit about their model and the hypothesized dynamics and mechanisms underlying their data, which gives readers the chance to critique the model and its hypotheses. The first step for the researcher is therefore to build a diagram that expresses the most plausible causal web for the variables of interest based on existing knowledge, theory, and evidence (Westreich & Greenland, 2013). One concern might be that life is too complex to be represented in one simple diagram, as there is always the possibility of unknown confounders, for example. But no one said research is easy. Nor should it be a mechanistic process of thoughtlessly adding more and more variables to an omnibus regression model. Models should be expressed transparently to allow examination and knowledge accumulation, as "rounds of model criticism and revision embody the real tests of scientific hypotheses" (McElreath, 2020, p. 139).

When it comes to ID research in language learning, correlational results abound. There is exploratory research where fad constructs are thrown into an analysis, theory-free, and are used to "explain" or "predict" a certain outcome. These "garbage-can regressions" (Achen, 2005) have little theoretical and causal rationale, make little empirical sense, and consequently provide no straightforward interpretation. With very few exceptions (e.g., Gardner, 2000), ID researchers also tend to avoid the topic of causality and its implications, sometimes treating it as an afterthought. For example, more than ten years after the L2 motivational system was proposed (Dörnyei, 2005), You et al. (2016) acknowledged that "the L2 motivational self system was originally proposed as a framework with no directional links among the three components" (p. 97). This situation led to contradictory conceptualizations even in chapters in the same edited volume (see Hiver & Al-Hoorie, 2020a). This situation also led to little accumulation of knowledge in the field over the past six decades, as well as little attention to coming closer to a consensus about a clear set of collective and measurable goals the field aspires to reach (Al-Hoorie et al., 2021). This is perhaps especially problematic for an "applied" field that purports to inform society. In short, a researcher needs a clearly articulated theory that explains what should be included and controlled for, and what should not, and why. It should no longer be tenable to justify the inclusion of a variable by lazy argumentation like "we know little about this variable, so we add it to the analysis."

In order to illustrate the concepts in this article, we present two examples next, drawing from *ability beliefs*. For the first example, consider the DAG in Figure 9. Just like other examples presented above, we use the model in Figure 9 as a methodological exercise (if other researchers argue for a modified model configuration, the principles will remain the same). In this model, the researcher is interested in the causal effect of teacher support on student achievement. Some variables in this model must be controlled, and others must not be controlled. The independent-dependent relationship in this model is confounded by assessment type, in that the type of assessment may have a washback effect on the level of teacher support as well as on student achievement. Therefore, assessment type must be controlled. If the researcher has not collected data on assessment type, this confounder must still be represented in the DAG – along with any other confounders known from theory – and the implications for the imprecision of the causal estimate must be discussed. Without this transparency, the researcher may simply advance generic claims that teacher support is a good "predictor" of student achievement.

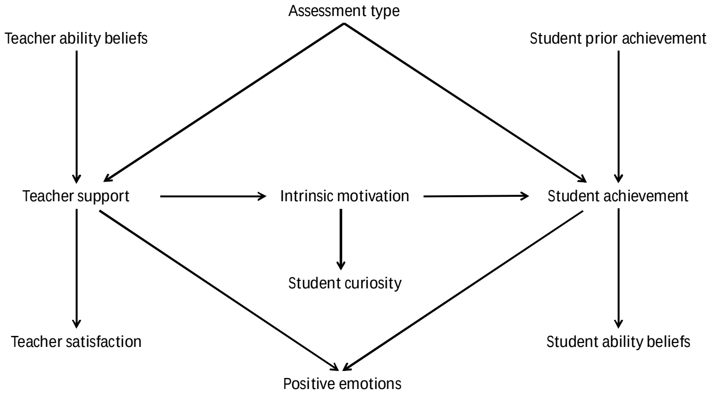

Figure 9 Example of a causal web

On the other hand, Figure 9 indicates that the researcher must not control for positive emotions. The same applies to similar constructs, such as student course satisfaction, the L2 learning experience, and attitudes toward the learning situation, if these are descendants of both teacher support and student achievement. Including these colliders in the model opens up a back-door path and introduces a non-causal association. The field has taken for granted that these variables exert a causal effect on student achievement (Al-Hoorie, 2018) despite evidence pointing to the contrary: "Students who expect to receive a good grade evaluate a teacher more positively than students who expect to receive a poor grade" (Stroebe, 2020, p. 227). This raises serious questions about the claim that the L2 learning experience is the "best predictor" of learning outcomes (for a discussion, see Al-Hoorie, 2017, pp. 160-166).

As for the other variables in the model, most must not be controlled (see Cinelli et al., 2022). Both intrinsic motivation and its descendant curiosity must not be controlled since this will remove the effect the researcher is interested in. An exception would be applying the front-door criterion if its assumptions are met as discussed above. Similarly, it would be a bad idea to control for teacher ability beliefs or teacher satisfaction, as this will reduce variation in teacher support (and, in the case of teacher ability beliefs, it may even lead to bias amplification). One situation where it might be useful to control for teacher ability beliefs is if it can be argued that it satisfies the assumptions of an instrumental variable as explained previously. Controlling for student ability beliefs is also harmful if it is a descendant of the dependent variable (also called a descendant of a virtual collider). Some researchers might argue that the variable "student ability beliefs" is not a descendant of achievement but a mediator between teacher support and achievement, but this still renders it identical to intrinsic motivation where controlling for it removes the intended effect in this particular model. The only variable that the researcher may control for is prior student achievement, which will improve the precision of the estimate.

This example shows that many types of variables are harmful if included as predictors. As explained above, researchers may argue for a different model configuration, and this is fine as long as the modified model is presented transparently and defended based on theoretical and empirical grounds. Variables should, of course, not simply find themselves in a model for random reasons to satisfy a capricious researcher. The same *d*-separation principles will apply to this modified, or improved, model. And in fact, this is how scientific knowledge accumulates incrementally. However, following conventional wisdom, the researcher may be tempted to simply use all of these variables, along with a couple more bandwagon variables-du-jour, as "predictors." This will arguably make the results meaningless.

In the second example, imagine that a researcher is now interested in finding out the contribution of ability beliefs and anxiety to student achievement (see Figure 10). There are different modeling possibilities for the relationships among these three variables, and each model requires a different data analytic approach. The relationship could be that of full mediation (see Figure 10, panel a) or partial mediation (see Figure 10, panel b). In these two cases, the researcher must not control for the mediator, unless it is explicitly modeled as a mediator to compute direct and indirect effects – assuming relevant assumptions hold. According to Figure 10, panel c, the relationship is confounded, and therefore the researcher must control for the confounder in order to obtain an accurate estimate. According to Figure 10, panel d, however, the relationship contains a collider, and so the researcher must not control for that collider to avoid opening a back-door path. Thus, with an apparently simple three-variable model, a number of possibilities emerge, each with a very different substantive interpretation. The researcher must first examine available theory and evidence, construct an explicit DAG of the relationships among the variables, defend these relationships, and then use the appropriate analytic procedures.

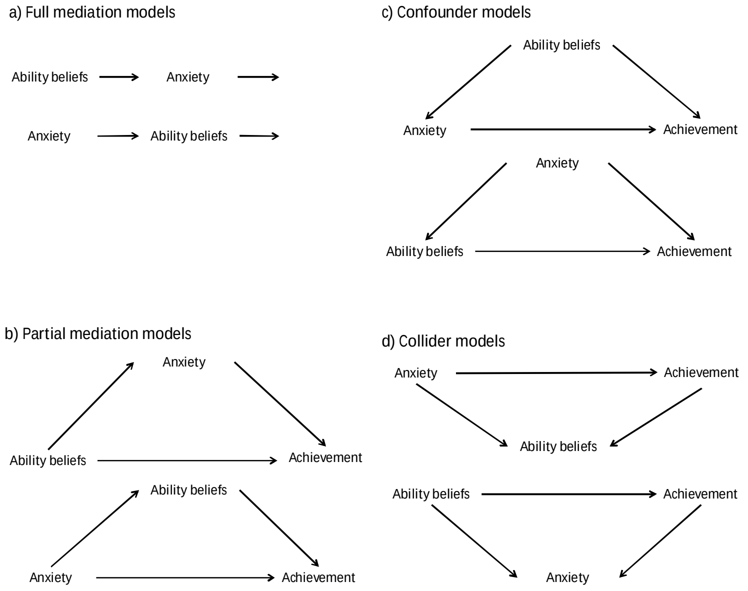

Figure 10 a) full mediation models; b) partial mediation models; c) confounder models, d) collider models

To summarize, researchers interested in incorporating the ideas presented in this article should follow three steps. In the first step, the researcher should think causally, not merely correlationally, at the study design stage. They should draw from theory, evidence, and experience to create and share a DAG representing the causal web of all variables relevant to their investigation to the best of their ability, including variables that will not be assessed in the study. The DAG may be drawn by hand or using specialized software. In the second step, the researcher should consider whether any variables, again whether assessed or not, might function as a confounder – or a common cause of the independent and dependent variables – or a descendant of a confounder. Such variables need to be included in the model and controlled for. If this is not possible, it still needs to be explicitly acknowledged as a limitation to causal inference resulting from this study design. In the third step, the researcher should consider whether any of the variables included in the study design function as a collider, or a common effect of both the independent and dependent variables. This includes descendants of colliders, as well as virtual colliders (descendants of the dependent variable). These variables must not be controlled for in the study design, whether through sample selection or stratification. Additionally, the researcher should consider whether any variable is a mediator or a descendant of a mediator and avoid controlling for it in a way that blocks the causal path. This level of causal transparency helps future research build on findings resulting from this study design cumulatively and meaningfully, and without reliance on mere correlations. One implication of this approach is that once a third variable (e.g., aptitude, prior ability) is established as a confounder in the relationship between an independent variable (e.g., motivation) and a dependent variable (e.g., achievement), the causal inference from any study that does not control for that confounder becomes suspect. Future research should therefore include such confounders to obtain a more precise estimate of the causal relationship.

5. Conclusion

A primary aim of this article is to call for a move in ID in language learning research from a thoughtless kitchen-sink approach to an informed variable selection process. In line with the principles of open scholarship, this variable selection process should be transparently communicated to the reader, preferably with a visual DAG. This cannot be substituted with replication, as (direct) replication is not designed to address conceptual ambiguities – and if anything, successful replicability using flawed designs may give the illusion that the results are valid (Al-Hoorie et al., 2024). On the other hand, conceptual replication, that intentionally re-examines design and variable selection, can address these ambiguities, especially when taking into account the level of theoretical maturity in the area of investigation (see substantiation framework; Al-Hoorie, Hiver, et al., 2023).

This article has also argued that similar problems arise through sampling procedures. Endogenous selection bias occurs when the researcher, intentionally or unintentionally, samples only a subset of the population before conducting the study, or loses participants (e.g., exclusion, noncompliance, or attrition) after conducting the study. Such preferential selection can introduce collider bias because the researcher, in effect, could be conditioning on (i.e., controlling for) an additional variable. This is a tricky situation because there is no explicit variable that the researcher is controlling for in the analysis, but it is conditioned on by design. Statistics cannot fix this problem. An example is selecting participants based on age or location (e.g., exclusively college students or WEIRD language learners) when the phenomenon under investigation is relevant to a broader section of the population. Procedures like multi-site replications cannot alleviate this problem if the participants sampled still do not represent the full range of the intended population. Thus, a multi-site replication may show that an effect replicates consistently, but this finding may still lack internal validity (flawed design) as well as external validity (not generalizable to the intended population).

Bigger is not always better. Nor does explaining more of the variance necessarily mean a better model. Stripping causal inference from research findings reduces them to mere mindless correlation-based predictions. In a causal web, there are typically many more (spurious) correlations than there are genuine causal relationships. Science progresses through thoughtful reflection on, critique of, and building on existing knowledge, whether experimentally or observationally, rather than through blind accumulation of correlations. We call on our colleagues and fellow researchers investigating IDs in language learning to take stock of available evidence-based knowledge, to eschew bandwagony variable-of-the-day approaches that have proliferated the field, and to return to the basic yet difficult work of explicitly modeling causal relationships through theoretically sound and careful empirical analysis.

References

Acharya, A., Blackwell, M., & Sen, M. (2016). Explaining causal findings without bias: Detecting and assessing direct effects. *American Political Science Review*, *110*(3), 512-529. https://doi.org/10.1017/s0003055416000216

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327-339. https://doi.org/10.1080/07388940500339167

Al-Hoorie, A. H. (2017). *Implicit attitudes in language learning* [Doctoral dissertation, University of Nottingham].

Al-Hoorie, A. H. (2018). The L2 motivational self system: A meta-analysis. *Studies in Second Language Learning and Teaching*, *8*(4), 721-754. https://doi.org/10.14746/ssllt.2018.8.4.2

Al-Hoorie, A. H., Hiver, P., & In'nami, Y. (2024). The validation crisis in the L2 motivational self system tradition. *Studies in Second Language Acquisition*, *46*(2), 307-329. https://doi.org/10.1017/S0272263123000487

Al-Hoorie, A. H., Hiver, P., Kim, T.-Y., & De Costa, P. I. (2021). The identity crisis in language motivation research. *Journal of Language and Social Psychology*, *40*(1), 136-153. https://doi.org/10.1177/0261927x20964507

Al-Hoorie, A. H., Hiver, P., Larsen-Freeman, D., & Lowie, W. (2023). From replication to substantiation: A complexity theory perspective. *Language Teaching*, *56*(2), 276-291. https://doi.org/10.1017/s0261444821000409

Al-Hoorie, A. H., Oga-Baldwin, W. L. Q., Hiver, P., & Vitta, J. P. (2022). Self-determination mini-theories in second language learning: A systematic review of three decades of research. *Language Teaching Research*, *29*(4), 1603-1638. https://doi.org/10.1177/13621688221102686

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444-455. https://doi.org/10.1080/01621459.1996.10476902

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

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552

Diener, E., Northcott, R., Zyphur, M. J., & West, S. G. (2022). Beyond experiments. *Perspectives on Psychological Science*, *17*(4), 1101-1119. https://doi.org/10.1177/17456916211037670

Ding, P., & Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*, *3*(1), 41-57. https://doi.org/10.1515/jci-2013-0021

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Lawrence Erlbaum.

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9-42). Multilingual Matters.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. Routledge.

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, *40*(1), 31-53. https://doi.org/10.1146/annurev-soc-071913-043455

Fiedler, K. (2011). Voodoo correlations are everywhere – not only in neuroscience. *Perspectives on Psychological Science*, *6*(2), 163-171. https://doi.org/10.1177/1745691611400237

Gardner, R. C. (2000). Correlation, causation, motivation, and second language acquisition. *Canadian Psychology/Psychologie canadienne*, *41*(1), 10-24. https://doi.org/10.1037/h0086854

Glynn, A. N., & Kashin, K. (2018). Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, *113*(523), 1040-1049. https://doi.org/10.1080/01621459.2017.1398657

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243-1255. https://doi.org/10.1177/1745691620921521

Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., Gillespie, M., Pazderka-Robinson, H., & Boadu, K. (2003). Pearl's d-separation: One more step Into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 289-311. https://doi.org/10.1207/s15328007sem1002_8

Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, *108*(5), 616-619. https://doi.org/10.2105/ajph.2018.304337

Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, *32*(1), 42-49. https://doi.org/10.1080/09332480.2019.1579578

Hiver, P., & Al-Hoorie, A. H. (2020a). Reexamining the role of vision in second language motivation: A preregistered conceptual replication of You, Dörnyei, and Csizér (2016). *Language Learning*, *70*(1), 48-102. https://doi.org/10.1111/lang.12371

Hiver, P., & Al-Hoorie, A. H. (2020b). *Research methods for complexity theory in applied linguistics*. Multilingual Matters.

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2021). Engagement in language learning: A systematic review of 20 years of research methods and definitions.

Language Teaching Research, *28*(1), 201-230. https://doi.org/10.1177/13621688211001289

Killingsworth, M. A. (2021). Experienced well-being rises with income, even above $75,000 per year. *Proceedings of the National Academy of Sciences*, *118*(4), e2016976118. https://doi.org/10.1073/pnas.2016976118

Lamb, M. (2017). The motivational dimension of language teaching. *Language Teaching*, *50*(3), 301-346. https://doi.org/10.1017/s0261444817000088

MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review*, *19*(1), 30-43. https://doi.org/10.1177/1088868314542878

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC Press.

Meyer, J., Lüdtke, O., Schmidt, F. T. C., Fleckenstein, J., Trautwein, U., & Köller, O. (2022). Conscientiousness and cognitive ability as predictors of academic achievement: Evidence of synergistic effects from integrative data analysis. *European Journal of Personality*. https://doi.org/10.1177/08902070221127065

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on post-treatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760-775. https://doi.org/10.1111/ajps.12357

Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238-247. https://doi.org/10.1177/2515245920917961

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688. https://doi.org/10.1093/biomet/82.4.669

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

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

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27-42. https://doi.org/10.1177/2515245917745629

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, *14*(2), 287-305. https://doi.org/10.1177/1094428110369842

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, *42*(4), 276-294. https://doi.org/10.1080/01973533.2020.1756817

Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2020). Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: Review and recommendations. *International Journal of Epidemiology*, *50*(2), 620-632. https://doi.org/10.1093/ije/dyaa213

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., & Ellison, G. T. H. (2016). Robust causal inference using directed acyclic graphs: The R package "dagitty." *International Journal of Epidemiology*, *45*(6), 1887-1894. https://doi.org/10.1093/ije/dyw341

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., Vehtari, A., & Hsu, D. (2021). *The piranha problem: Large effects swimming in a small pond*. arXiv. https://doi.org/10.48550/arXiv.2105.13445

Westreich, D., & Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292-298. https://doi.org/10.1093/aje/kws412

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science*, *5*(2), 1-19. https://doi.org/10.1177/25152459221095823

You, C., Dörnyei, Z., & Csizér, K. (2016). Motivation, vision, and gender: A survey of learners of English in China. *Language Learning*, *66*(1), 94-123. https://doi.org/10.1111/lang.12140

Young, S. S., & Karr, A. (2011). Deming, data and observational studies. *Significance*, *8*(3), 116-120. https://doi.org/10.1111/j.1740-9713.2011.00506.x

Zhou, S., Hiver, P., & Al-Hoorie, A. H. (2021). Measuring L2 engagement: A review of issues and applications. In P. Hiver, A. H. Al-Hoorie, & S. Mercer (Eds.), *Student engagement in the language classroom* (pp. 75-98). Multilingual Matters.