METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

How Large Must an Associational Mean Difference Be to Support a Causal Effect?

Michael Höfler[1,2], Ekaterina Pronizius[3,4], Erin Buchanan[5]

[1] *Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany.* **[2]** *Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany.* **[3]** *Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria.* **[4]** *Faculty of Psychology and Educational Sciences, Psychological Sciences Research Institute, University of Louvain, Louvain-la-Neuve, Belgium.* **[5]** *Analytics, Harrisburg University of Science and Technology, Harrisburg, PA, USA.*

Methodology, 2024, Vol. 20(4), 318–335, https://doi.org/10.5964/meth.14579

Received: 2024-05-07 • **Accepted:** 2024-11-29 • **Published (VoR):** 2024-12-23

Handling Editor: Jochen Mayerl, Chemnitz University of Technology, Chemnitz, Germany

Corresponding Author: Michael Höfler, Chemnitzer Straße 46, Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, 01187 Dresden, Germany. +49 351 463 36921. E-mail: michael.hoefler@tu-dresden.de

Supplementary Materials: Code, Materials [see Index of Supplementary Materials]

Abstract

An observational study might support a causal claim if the association found cannot be explained by bias due to unconsidered confounders. This bias depends on how strongly the common predisposition, a summary of unconsidered confounders, is related to the factor and the outcome. For a positive effect to be supported, the product of these two relations must be smaller than the left boundary of the confidence interval for, e.g., a standardised mean difference (*d*). We suggest means to derive heuristics for how large this product must be to serve as a confirmatory threshold. We also provide non-technical, *visual means* to express researchers' assumptions on the two relations to assess whether a finding on *d* is explainable by omitted confounders. The *ViSe* tool, available as an *R* package and Shiny application, allows users to choose between various effect sizes and apply it to their own data or published summary results.

Keywords

causality, confirmation, observational studies, effect size, visualisation, software

Psychologists usually avoid appealing to causality outside of experimental studies (Rohrer, 2018). At the same time, associations in observational studies are often reported as if they were plausible proxies for causal effects, which gives them more meaning (Alvarez-Vargas et al., 2023; Grosz et al., 2020). Causal meaning is then denied again when, after conducting an observational study, authors remind the readership that 'correlation is not causation' (Pearl & Mackenzie, 2018). Through this assertion, readers are urged to be cautious when a factor of interest, such as child maltreatment, cannot be subjected to manipulation or randomization. Yet the causal relations that form theories must be investigated (Hernán, 2018), as does the potential of interventions (Pearl & Mackenzie, 2018).

We argue that associations can still be informative for causal analysis if one predicts an association of a certain size that allows for a reasonable amount of bias. Then a study is conducted to test whether the association exceeds that size. In contrast, just taking the association as a proxy means testing only that it is greater than zero, which may constitute a poor test that undermines the hypothesis's falsifiability. The critical question is how *large* the association needs to be to *withstand bias*.

We focus on bias due to confounders (common causes), which is the specific issue of causal inference (Pearl, 2009). Such bias depends on the strength of the confounders' influence on both the factor and the outcome, for example, when estimating the effect of child maltreatment on adulthood internalising problems (Kisely et al., 2018). Researchers can adjust for some confounders to remove bias from factor-outcome associations; however, unconsidered confounders may still introduce bias (Hernan & Robins, 2020). An extensive toolbox of formal methods has been developed to identify which confounders to consider and how exactly to do adjustment in analysis (Hernan & Robins, 2020), but has been underutilised in some fields including psychology.

In this paper, we propose a *visual approach* that offers a more accessible alternative for researchers who find these methods challenging. After discussing simple means for deriving heuristics on the amount of bias that an association must exceed, we present *ViSe*, a graphical tool, implemented as *R* package and Shiny app. The tool allows researchers to illustrate under which assumptions on unconsidered confounders a causal effect is supported (Buchanan, 2024). Our focus is on associations measured by the mean difference between two groups, or effect size ***d***. However, our approach can also be applied to other effect sizes, as the tool facilitates their conversion into ***d***. Note that we do not address within-group effects, since these differ both technically and conceptually (confounders that remain constant over time are inherently accounted for by design).

Method

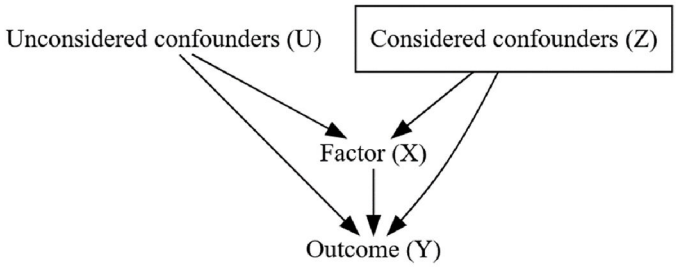
Generative Model

Data do not 'speak for themselves'. Between the causal effects operating in reality and the observable data lie processes such as confounding, selection, and measurement. Relying solely on data presumes that these processes either do not exist or do not influence the estimate of an effect. Instead of relying on these undefendable assumptions, one can model these processes, which justifies a causal interpretation, provided the model holds (Greenland, 2005). 'Model' means firstly that one builds a *generative model* of these processes. We use the common method of directed acyclic graphs (DAGs) for this. DAGs are based on simple qualitative assumptions of which variables influence both factor and outcome and what causal relations between them exist, represented with arrows. No assumptions on the size and functional form of the effects and on the distribution of the variables are necessary (Torchiano, 2020).

Figure 1 shows our DAG model. Its assumptions on confounders might apply to many effects in clinical psychology. The model supposes that a *common predisposition*, summarised in a single variable (**U**) and labelled 'unconsidered confounders', affects both factor (**X**) and outcome (**Y**). The factor-outcome association may already be adjusted for some variables, 'considered confounders' also summarised in a variable (**Z**).

Figure 1

*A Generative Directed Acyclic Graphs Model Illustrating the Effect of a Factor (**X**) on an Outcome (**Y**)*



Note. We use the example of child neglect (**X**) on internalising problems in adulthood (**Y**). Socio-demographic variables (**Z**) have been adjusted for, as indicated by the box around **Z**. Other confounders including parental and bonding characteristics (**U**) may affect both **X** and **Y**. If these cannot be explained by **Z**, bias arises in estimating the effect of **X** on **Y**. Arrows represent causal relations.

Other variables may also influence the outcome, but they can be omitted from the DAG if they do not also influence the factor. *DAGitty* (Textor et al., 2016) is a valuable *R* package and Shiny app for graphing and evaluating DAGs, such as identifying which variables

require adjustment. In addition, it allows one to assess the compatibility of a DAG with the data by showing all the associations predicted by the model (Glymour et al., 2019).

Most importantly, *one cannot have no generative model*. In observational studies, a factor of interest, like child maltreatment, is influenced by its natural causes, including parental and bonding factors, which may also impact the outcome, e.g., adult mental health problems. Not adjusting for confounders implicitly assumes that none exist. If this assumption holds, Figure 1 would exclude both 'considered' and 'unconsidered' confounders. Considering only unspecific variables, like sex and age, as common, supposes that child maltreatment and later mental health issues share no other causes beyond these variables—thus eliminating 'unconsidered confounders' from Figure 1. Such assumptions lack transparency, may go unrecognized, and are generally undefendable, undermining effective scientific communication (Deffner et al., 2022).

In epidemiology, various methods of 'sensitivity analysis' have been developed to unravel the impact of quantitative assumptions on the results of causal analyses and require at least basic mathematical understanding (D'Agostino McGowan, 2022; Fox et al., 2021). Most of these methods focus on binary outcomes and effects measured on the risk or odds ratio scale. Here we present, a) a means of *deriving simple heuristics* for determining how large an association must be, and b) a completely non-technical, *visual approach*. Our focus is on interval-scaled outcomes, which are common in clinical psychology where non-randomized groups are compared on mental health measures. We also assume that the common disposition (**U**) is an interval-scaled continuum.

Variable Omission Model

Before we can quantify and visually assess the impact of assumptions, we must complete the model, the nonparametric DAG from Figure 1, with *quantitative (parametric) assumptions*. For an interval-scaled outcome (**Y**), the following 'variable omission model' has been widely applied in the literature (Cinelli & Hazlett, 2020; D'Agostino McGowan, 2022; Gelman & Hill, 2006). This model suggests that bias in estimating the factor-outcome (**X**–**Y**) effect arises from omitting unconsidered confounders (**U**) in the linear regression of **Y** on dummy-coded **X**, possibly adjusting for some confounders **Z**. The model relies on three regression equations:

1. **Y** ~ **X** + **U** (+ **Z**)
2. **Y** ~ **X** (+ **Z**)
3. **U** ~ **X**

The bias in estimating the effect of factor **X** on outcome **Y** equals simply the product $\beta_{UY} \times \beta_{XU}$, where β_{UY} is the regression coefficient of **Y** on **U**, and β_{XU} is the regression coefficient of **U** on **X**. Thus, bias depends on two factors: the effect of unconsidered confounders **U** on, 1) **X** and 2) **Y**. Since **U** is a latent variable with an arbitrary scale, we may assume it to be standardised (variance = 1). Also, the observed outcome **Y** may

be scaled in standard deviations, in which case the associational mean difference is a standardised mean difference, $\boldsymbol{d_{XY}}$. (Note that, we omit the adjustment on **Z**; for example, $\boldsymbol{d_{XY}}$ might actually imply $\boldsymbol{d_{XY|Z}}$). For ease of notation, we refer to it as ***d***, a member of the '*d* Family', which comprises effect sizes that 'express the mean difference in standard deviation units', but differ in how the standard deviation used for standardisation is computed, yielding Cohen's *d*, Glass's delta or Hedge's *g* (Kraemer et al., 2003).

Now, the bias in the estimated causal effect of factor **X** on outcome **Y** on the standardised mean scale, ***d***, through disregarding the confounders **U,** can be written as:

$$(*)\ \mathbf{c} = \boldsymbol{d_{XU}} * \boldsymbol{corr_{UY}}.$$

$\boldsymbol{d_{XU}}$ is the confounder-factor **effect** on the Cohen's *d* scale (i.e., by how many standard deviations does outcome **Y** differ between factor groups **X** = 0 and **X** = 1?), and $\boldsymbol{corr_{UY}}$ is the confounder-outcome effect on the standardised scale (i.e., if the confounder summary **U** increases by one standard deviation, by how many standard deviations does outcome **Y** change?). We abbreviate it as *'corr'*, because it is scaled like a correlation.

Precisely, the variable omission model adds the following parametric assumptions to the DAG:

1. The confounder summary **U** is normally distributed (or can be imagined to be transformed to a normal distribution) in both factor groups **X** = 0 and **X** = 1 with equal variances (set to 1, perhaps after adjusting for **Z**). (Regarding **U** as 'common predisposition' stresses that the variables contributing to **U** must precede **X**, because otherwise they could be mediators of the effect of **X** on **Y**, in which case they must not be adjusted for).
2. In both factor groups, **X** = 0 and **X** = 1, the confounder **U** and the outcome **Y** are linearly related with equal slope.

$\boldsymbol{d_{UX}}$ describes how much the distribution of confounder **U** in **X** = 1 (e.g., individuals with child maltreatment) is shifted as compared to **X** = 0 (individuals without child maltreatment). The smaller the difference, the better the groups are *balanced* in analogy to randomised studies, where $\boldsymbol{d_{UX}}$ can only differ from 0 by chance, and thus, is expected to equal 0 (Cinelli & Hazlett, 2020).

Illustrative Explanation

Suppose that factor **X** has a positive effect on outcome **Y**. The relation (*) then essentially says that the associational mean difference ***d*** must *exceed* $\mathbf{c} = \boldsymbol{d_{XU}} * \boldsymbol{corr_{UY}}$ to allow for bias due to unconsidered confounders (in case of a negative effect it must be *less* than **c**). To also account for random error, the *left boundary* of the confidence interval for ***d*** must exceed **c** (see below).

In summary, *a high estimated* $\boldsymbol{d}$ together with a *small assumed* $\boldsymbol{d_{XU}}$ or $\boldsymbol{corr_{UY}}$ supports a causal effect. If this is the case, an association between **X** and **Y** cannot be explained by common causes. In particular, the simple relation (*) has the following implications:

- Unconsidered variables (**U**) must affect *both* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ to produce bias $\mathbf{c} \neq 0$.
- Bias **c** is large only if *both* $\boldsymbol{d_{XU}}$ *and* $\boldsymbol{corr_{UY}}$ are large, thus *unconsidered confounders have large effects on both factor and outcome*. The meaning of 'large' is initially vague, but will become clearer the more researchers are enabled to assess these quantities.
- If $\boldsymbol{d_{XU}}$ and $\boldsymbol{corr_{UY}}$ have the same sign, that is, the effects of unconsidered variables on factor and outcome are *both positive* or *both negative*, bias is upward, $\mathbf{c} > 0$; if they have different signs, bias is downward, $\mathbf{c} < 0$.

Applied to our example, it has been hypothesised that poor parenting, mental disorders, adverse life events, poor social skills, and other unfavourable or detrimental conditions are generally *positively relate*d (Höfler et al., 2021). Under this assumption, we would expect $\mathbf{c} > 0$ for effects, for instance, of child maltreatment on later mental health problems.

Determination and Heuristics on *c*

(*) provides a simple formula to calculate **c**. We conceptually refer to it as the *confirmation threshold*. Theoretical considerations may guide the specification of its factors $\boldsymbol{d_{UX}}$ and $\boldsymbol{corr_{UY}}$. In the example, the groups may be poorly balanced beyond sociodemographic variables, with greater exposure to adverse conditions, such as parental and parenting factors, among those experiencing maltreatment. This might suggest a medium $\boldsymbol{d_{UX}}$ around 0.65, which is the average of the typical range for medium effect sizes (0.5–0.8). Adverse conditions also affect internalising and externalising problems, but since these issues develop later, $\boldsymbol{corr_{UY}}$ might fall within the small to medium range—around 0.20, according to common standard. Combined, **c** might be chosen around 0.13.

In the absence of such a specific theoretical foundation, one might refer to the hypothesis in behavioural science that 'everything correlates with everything' due to a vague underlying multivariate causal structure (Orben & Lakens, 2020). Their literature review suggests that a correlation of 0.30—seen for variables occurring not far apart in time—is the highest plausible value. Thus, this value can serve as a conservative choice. According to (*), causal effects may be smaller than correlations, particularly if all undesired conditions are positively related (Höfler et al., 2021). If three variables are pairwise correlated at 0.30, the partial correlation between any two variables, adjusting for the third, is 0.23. Assuming that this value is roughly applicable to the proximal predisposition-factor (**U**–**X**) relation, converting to *d* yields $\boldsymbol{d_{UX}} = 0.33$. Now, the causal correlation for the more distal predisposition-outcome (**U**–**Y**) relation might be smaller than 0.23, perhaps by ⅓, ½, or ⅔, giving $\boldsymbol{corr_{UY}} = 0.15$, 0.12 and 0.08, respectively (conservatively rounded). Thus, without specific theoretical assumptions, a crude heuristical

c might equal 0.05, 0.04 or 0.03, giving a conservative choice of 0.05. Importantly, due to the weak substantive foundation, this choice is not warranted if there is substantive grounds for higher thresholds. Uncertainty in the existence of confounders with strong effects on factor and outcome does not justify acting as if they definitely do not exist. We generally suggest a conservative approach in the spirit of severe testing (Lakens et al., 2024), choosing the highest plausible **c**.

Hypothesising a Causal Effect May Predict an Association

Epistemically, in the Popperian tradition of science (Lakens et al., 2024; Popper, 2002), a hypothesis about an effect predicts an association of a certain magnitude if the presumed model assumptions (here the DAG and the variable omission model with the two specified bias quantities) hold. The hypothesis is then falsified if an association of that magnitude is not found (at least it is not confirmed). The notion of magnitude here is akin to Bradford Hill's (1965) consideration that the larger the association, the more likely a causal effect is.

How to Obtain *d* and Statistical Inference On It

Before computing what we will call the 'sensitivity plot' to visualise whether a causal effect is supported, we need to mention some technical issues about $\boldsymbol{d}$. Researchers typically aim to show that the factor-outcome association is greater than 0 through testing H_0: $\boldsymbol{d} \leq 0$ against H_1: $\boldsymbol{d} > 0$. This test is equivalent to the *left bound* of the one-tailed $1 - \alpha$ confidence interval, $\mathbf{l}_\alpha$ (shortly $\boldsymbol{l}$) exceeding 0. (For alleged negative effects, H_0: $\boldsymbol{d} \geq 0$ against H_1: $\boldsymbol{d} < 0$, and the right bound of the one-tailed $1 - \alpha$ confidence interval must be less than 0. The implementation in *ViSe* allows changing the sign of $\boldsymbol{d}$ in the hypothesis).

While we focus on one-tailed tests, our approach and its implementation can also be using two-tails, which applies if the direction of the association has not been predicted. In this case, $\boldsymbol{d}$ has to be replaced with its absolute value, $|\boldsymbol{d}|$, and $\mathbf{l}_\alpha$ with $\mathbf{l}_{\alpha/2}$; thus $|\boldsymbol{d}|$ or $\mathbf{l}_{\alpha/2}$ can be entered into the computation of a sensitivity plot in *ViSe*. In the absence of confounder adjustment, the left bound **l** can be taken from an analysis that calculates $\boldsymbol{d}$ and a confidence interval for it, for example with the *R* package *effsize* (Torchiano, 2020).

In case of adjustment, linear regression of the outcome on the factor (dummy-coded) and the considered confounders is often used. Robust linear regression accounts for violated assumptions (non-normal distributions, extreme values, unequal variance; Mair & Wilcox, 2020). If the outcome is standardised beforehand (divided by its standard deviation), the estimate of the $\boldsymbol{\beta}$ for the factor-outcome relation can be interpreted on the $\boldsymbol{d}$-scale. Such an estimate is (noncentrally) *t*-distributed, and the left bound **l** is computed from that distribution, given standard error (SE) and degrees of freedom. In large samples, the distribution of an estimate of $\boldsymbol{\beta}$ may be approximated with a normal distribution (Severini, 2000). In this case, **l** is simply calculated as:

$$\mathbf{l} = \text{estimate} - \Phi(1-\alpha) * SE$$

Φ = cumulative of standard normal distribution, with $\alpha = .05$, $\Phi(0.95) = 1.64$.

ViSe provides both the normal and noncentral *t*-distribution estimates for the lower bound **l**, as well as a two-sided confidence interval if one wishes to be conservative. Now, to demonstrate a causal effect based on the above generative model, **l** > **c** must be found. Thus, the confirmation threshold (**c**) is the amount of correction in the associational test.

Empirical Anchoring to Determine the Confirmation Threshold

The above theoretical and heuristic considerations for determining **c** are *a priori*, that is, they must precede, or at least be independent of the data analysis. Another but *post hoc* method is worth mentioning. It makes use of the analysis' result, namely the *change* in the estimate of ***d*** through adjusting for the considered confounders. This change provides an anchor for assessing potential bias from other factors (Cinelli & Hazlett, 2020). The method reveals the maximum possible size of **c** relative to this change, but requires a large study with small random error. For example, if a study estimated the lower bound of ***d*** at 0.20, but adjustment for confounders reduced it to 0.15 (a change of 0.05), then unadjusted confounders would need to account for at least three times this change to nullify the result (0.20 - 0.05 - 0.15 = 0). However, if the adjustment reduced ***d*** to 0.05 (change = 0.15), unadjusted confounders would only need to account for one-third of the change (0.20 - 0.15 - 0.05 = 0). In the first case, one might accept the effect if theory suggests that unadjusted confounders are unlikely to dominate, while in the second case, one might reject it. In situations where anchoring is insufficient to determine **c**, it is necessary to evaluate the quantities affecting **c**, as we do visually below.

Testing for a Smallest Effect Size of Interest

If the goal is to show the effect exceeds not just 0 but a threshold Δ, e.g., the 'smallest effect size of interest' (Anvari & Lakens, 2021), the lower bound **l** must exceed **c** plus Δ. Adding Δ ensures the effect exceeds the minimum size of interest, commonly $\Delta = 0.2$. For proper confirmation, **c** (and Δ) must be specified a priori, which can be verified through pre-registration (Lakens et al., 2024). As shown in the online Supplement (Höfler et al., 2024), adding Δ may necessitate a much larger sample size.

Determination of the Quantities With Visual Aid

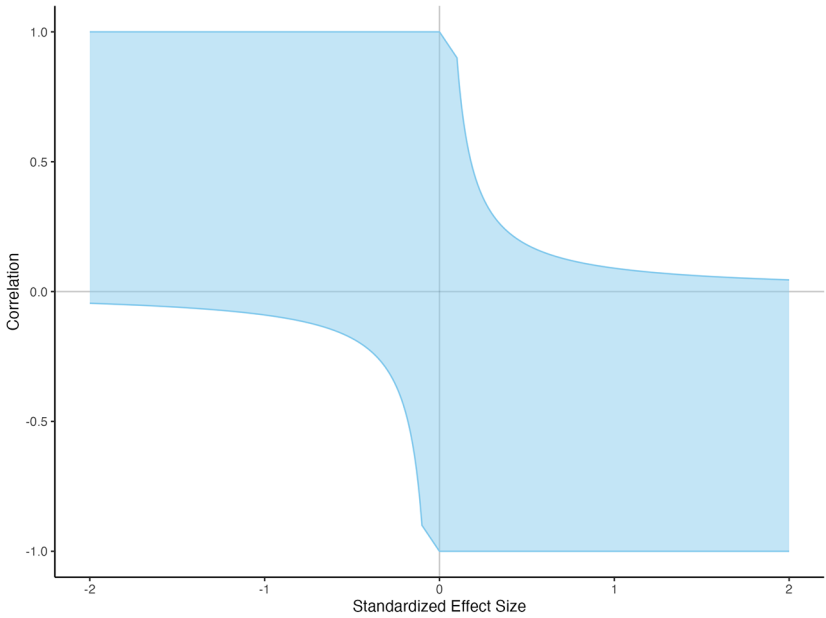
So far, we have used nonparametric (Figure 1) and parametric assumptions (variable omission model) to shed light on the bias due to unconsidered confounders. Still, how well correcting statistical inference by requiring the lower bound ***l*** to exceed **c** (+ Δ) rather than just 0 (Δ) works, depends on how accurately researchers can narrow down the two quantities that determine the confirmation threshold **c**.

We use *graphical means* for this, because visual perception of thoughtfully designed graphs facilitates researchers' choices and avoids the obstacles inherent in quantitative choices (e.g., Eberhard, 2023). Visual methods have also been proposed to support causal analysis (Guo et al., 2023). A *sensitivity plot* like the following post-hoc illustrates which combinations of the confounder-factor relation ($\boldsymbol{d_{XU}}$) and the confounder-outcome relation ($\boldsymbol{corr_{UY}}$) yield a lower bound (**l**) that exceeds the confirmation threshold (**c**). The plot uses **l** = 0.09 as found in the example (Kisely et al., 2018) that we will address in detail in the next chapter.

The plot in Figure 2 is enhanced by visual tools that assist in the specifications of the confounder-factor and the confounder-outcome relation. The specified values are mapped into the sensitivity plot; for example, 'based on your choices, **c** equals 0.2, and this value is (in)consistent with the causal hypothesis'.

Figure 2

Sensitivity Plot With $\boldsymbol{l}$ *= 0.09 and Two-Tailed* α *= .05*



Note. The shaded light blue area represents all combinations of the standardised effect size (x-axis) and correlation (y-axis) that support effect > 0 ($\mathbf{c} = \boldsymbol{d_{XU}}^* \boldsymbol{corr_{UY}} < \mathbf{l}$). The plot was created with the following specifications using the *ViSe* package: visualize_c(dlow = 0.09, lower = TRUE).

Specification of the Bias Quantities

In the example unconsidered confounders (**U**) summarise *parental and bonding characteristics*, a predisposition for both the factor child maltreatment and the outcome adulthood

internalising problems. The first quantity, $\boldsymbol{d_{XU}}$ (effect of the predisposition on child maltreatment), can be expressed and converted into different effect sizes. Providing multiple options helps to approximate its value and addresses the current lack of knowledge on which effect size works best for this.

1. $\boldsymbol{d_{XU}}$ can be directly visually represented by the *amount of shift* in the **U** distribution between the factor groups (**X** = 0 and **X** = 1). *This shows how much more disposition due to parental and bonding characteristics those with childhood maltreatment are assumed to have than those without.*
2. $\boldsymbol{d_{XU}}$ may be converted into the *proportion overlap* between the **U** distributions in **X** = 0 and **X** = 1, given by $2 * \phi(-1/2 * |d_{\boldsymbol{XU}}|)$, where ϕ is the cumulative of the standard normal distribution (Pastore & Calcagnì, 2019). This quantity equals *the fraction of the disposition that is believed to be shared by those with and without child maltreatment.* Researchers, however, might find other effect sizes more appealing: U1, the proportion of non-overlap across both distributions; U2, the proportion in one group that has higher scores than the same proportion in the other group; U3, the proportion of one group that is smaller than the median of the other group (Cohen, 1988).
3. *Superiority* (or 'common language effect size' or 'area under the ROC curve') is the probability that an individual in **X** = 1 scores higher in **U** than an individual in **X** = 0 and can be calculated as $\phi(|d_{\boldsymbol{XU}}| / \sqrt{2})$. (Furukawa & Leucht, 2011). Superiority indicates *how much more likely an individual with child maltreatment is to have a larger disposition than an individual without child maltreatment.*
4. $\boldsymbol{d_{XU}}$ can also be obtained by its relation with the *number needed to treat* (NNT), the number of individuals who, if there were truly a causal effect of **X** on **Y**, would (in expectation) have to undergo **X** = 1 in order to achieve a superior **U** than an individual with **X** = 0 (Furukawa & Leucht, 2011). $\text{NNT} = 1 / (2 * \phi(|d_{\boldsymbol{XU}}| / \sqrt{2} - 1))$. It indicates *how many more individuals maltreated as a child would cause one more individual with a larger disposition.*
5. Besides, $\boldsymbol{d_{XU}}$ can be converted into common variants, *f*, f^2, and *r* using $f = |d_{XU}| / 2$, and the approximation of *d* to *r* using $r = |d_{XU}| / \sqrt{(|d_{XU}|^2 + 4)}$. *R* can be interpreted similarly to a correlation, or alternatively the explained variance r^2 may be specified.
6. All of the above effect sizes can be displayed together to force a *coherent choice*. This is achieved by a 'donut chart', where the selection of one effect size automatically adjusts the circles representing the other effect sizes and their corresponding values (Calin-Jageman, 2018). For example, choosing $\boldsymbol{d_{XU}} = 0.5$ gives proportional overlap = 80%, superiority = 64%, and NNT = 3.62. Presenting these conversions side by side may help narrow down the quantity, as a value on one scale (e.g., 80% overlap) may seem small but is equivalent to a medium effect size of $\boldsymbol{d_{XU}} = 0.5$ by common standard.

The following Figure 3 shows the donut chart as implemented in *ViSe*:

Figure 3

Donut Chart Computed From ***d*** *= 0.50*

Note. Use visualize_effects(d = .50, circle_color = "blue", percent_color = "blue", text_color = "blue") in *ViSe* to create this chart.

Note that the conversion formulas above assume a normal distribution with equal variance, similar to the variable omission model. Each choice results in a different ($\boldsymbol{d_{XU}}$, $\boldsymbol{corr_{UY}}$) point on the sensitivity plot, which may or may not fall within the region confirming a causal effect. (Also note that through conversion into $\boldsymbol{d_{XU}}$, results reported on other effect sizes than $\boldsymbol{d_{XU}}$ can be handled with the methods described in this paper, although this may be a bit crude).

For specifying the second bias quantity, the confounder-outcome relation $\boldsymbol{corr_{UY}}$ (effect of the predisposition on adulthood internalising problems), there are fewer options, most importantly scatterplots. We use scatterplots with points that represent linear relations, i.e., slopes and equivalent correlations of varying magnitudes. In the example, the standardised slope indicates by how many standard deviations mental health problems in adulthood changes per increase in disposition due to parental and bonding factors by one standard deviation (in both of the groups). The square root of the standardised slope equals the variation in mental health problems explained by disposition (since $R^2 = \boldsymbol{corr_{UY}}^2$).

Finally note that all these *ViSe's* visualisation and conversion features can also be used *before* data collection to inform the choice of the confirmation threshold **c**.

Results

Example

We now illustrate in detail the use of the *ViSe* tool with the example of child maltreatment's effect on internalising behaviour at around age 21 (using the Youth Self-Report, YSR, scale). The study by Kisely et al. (2018) used a general population sample in Brisbane, Australia, comparing 3,554 mother-child pairs without 'substantiated child maltreatment' to 73 pairs with child neglect. We chose child neglect among other maltreatment that the study has investigated because of its suitability for demonstrating the tool. Maltreatment was assessed 'by linking to data from state child protection agencies'. The study reports unadjusted and adjusted mean differences, adjusting for sociodemographics including gender, parental ethnicity, and maternal education (likely using ordinary least squares regression, though not specified). We divided the reported mean differences by the standard deviations of the total sample to obtain the estimates and 95% confidence intervals on the ***d*** scale (via $\boldsymbol{d} = M_{difference} / SD_{Total}$) (see Table 1).

Table 1

Mean Outcome Values Reported in Table 1 (Kisely et al., 2018)

	Internalising score		
Difference	**Unadjusted**	**Adjusted**	**SD[b]**
Reported (raw difference)	3.68 (1.73–5.62)	2.73 (0.77–4.69)	8.29
Transformed[a] (standardised difference, d)	0.44 (0.20–0.68)	0.33 (0.09–0.57)	1

[a] Note that the boundaries of the confidence intervals are conservatively rounded. The left boundaries are rounded down; the right boundaries are rounded up (the point estimates are rounded numerically). [b] Standard deviation of the total sample.

The unadjusted ***d*** score equals 0.44 with a confidence interval of 0.20 (rounded down) to 0.68. After adjustment, ***d*** is lowered to 0.33 (0.09–0.57). The left bound **l** = 0.09 leaves some room for bias due to omitted confounders. While the theoretical confirmation threshold **c** = 0.13 from the last chapter is clearly not passed, the heuristic thresholds (0.04–0.06) would be. However, there is no evidence for an effect of sufficient size of interest with $\Delta = 0.20$ in any case.

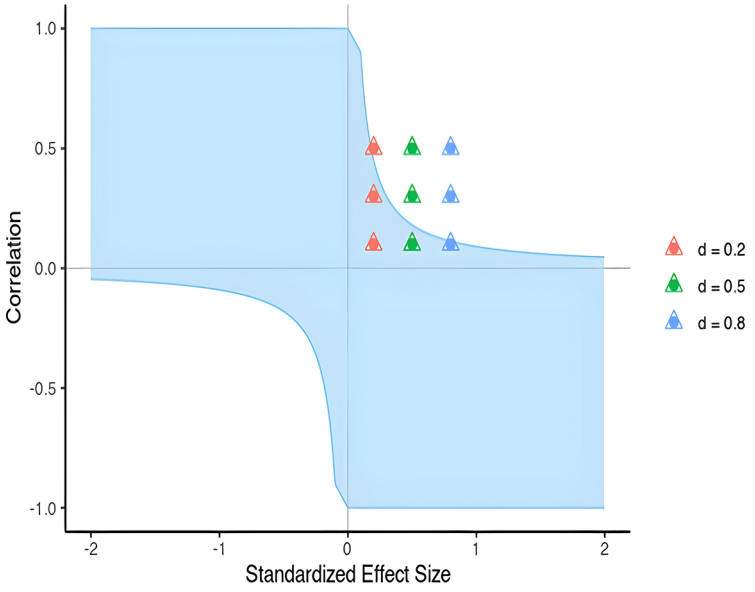
With regard to the anchoring approach to determine **c**, adjustment for socio-demographic variables reduced ***d*** by 0.11 (0.20–0.09). Thus, bias from unconsidered confound-

ers must be less than 82% of this change (0.09 / 0.11). Unconsidered confounders may include prior parental mental disorders, genetic factors, maltreatment history, stress, social support, attachment patterns, and parent-child relationship quality (Daníelsdóttir et al., 2024).

Now, what values of $\boldsymbol{d_{XU}}$ and $\boldsymbol{corr_{UY}}$ (or a conversion) would pass the confirmation threshold allowed by the result of l = 0.09? Figure 4 is a sensitivity plot with nine joint specifications of $\boldsymbol{d_{XU}}$ (0.2, 0.5, 0.8) and $\boldsymbol{corr_{UY}}$ (0.1, 0.3, 0.5). The plot covers wide ranges, allowing users to assess the plausibility of the displayed quantities. Four out of nine specifications [$\boldsymbol{d_{XU}}$ (0.2) & $\boldsymbol{corr_{UY}}$ (0.1, 0.3, 0.5) || $\boldsymbol{d_{XU}}$ (0.5) & $\boldsymbol{corr_{UY}}$(0.1)], indicating modest bias, support an effect of child neglect on adult internalising problems.

Figure 4

Sensitivity Plot for the Effect of Child Neglect on Adult Internalising Problems With Various Specifications for Bias Quantities d and r



Note. The plot was created using the *ViSe* package with the following settings: visualize_c_map(dlow = 0.09, r_values = c(.1, .3, .5), d_values = c(.2, .5, .8), lower = TRUE).

The Supplement (see Höfler et al., 2024) describes the implementation and use of *ViSe*. It provides further examples to illustrate the specification of the bias quantities.

Discussion

We outlined how to derive confirmation thresholds that cannot be explained by unconsidered confounders. This overcomes the common practice of only testing if an association exceeds 0. We also provided visual guidance on how to determine the necessary quantities: the effects of a common predisposition on both the factor and the outcome. Initially, when visual bias assessment is new to many researchers, solid examples are essential for training. Such examples can serve as anchors for specifying the two quantities and their product (extent of bias). To train their usage, general examples from previous causal (meta-)analyses (Mathur & VanderWeele, 2022) and specific cases where a causal effect is well understood should help. Well-understood examples enable empirical investigation into the most effective specification of the quantities, and which of the convertible effect sizes work best for this. An interesting application for better thresholds to suggest causality is network analysis to identify causal structures (Glymour et al., 2019). In their algorithms, the usual association threshold of 0 can be replaced by a defensible choice.

The proposed method for specifying bias quantities and graphically assessing their impact on causal inference has focused on binary factors, interval-scaled outcomes and the effect size ***d***. Possible extensions include:

1. Explicit assessment of **effect sizes** for non-standardised or binary outcomes and interval-scaled factors (rather than crude conversion into ***d***).
2. **Generative Models** that include other sources of bias, such as selection processes (Deffner et al., 2022) and measurement error (Fox et al., 2021), and jointly address **multiple bias** (Greenland, 2005). Together with quantitative assumptions, this would open the door for confirmation thresholds in more complex situations, which might turn out to be larger or smaller.

Note that selection can be partially addressed within the framework provided here. This holds if selection is related to the factor, but not to the outcome—beyond considered and unconsidered confounding factors. By studying a selective, homogeneous population, the relations of unconsidered confounders with factor and outcome ($\boldsymbol{d_{XU}}$ and $\boldsymbol{corr_{UY}}$) are expected to be *larger*, since homogeneity means smaller SDs in their denominators. Thus, the confirmation threshold must be set *higher*.

By initially focusing on a specific type of bias, namely confounding, we aim for our work to contribute to the establishment of stricter and more defensible confirmation thresholds in psychology and other disciplines (Höfler et al., 2022). Our basic approach involves specifying and multiplying only two quantities, a process that can be performed across scales and is supported by straightforward visual evaluation.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Acknowledgments: The authors thank Anna-Lena Zietlow for advice on common causes of neglect and internalising disorders and Julia Rohrer for carefully reviewing the manuscript.

Competing Interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplementary Materials

For this article, the following Supplementary Materials are available:

- R code. (Höfler et al., 2024)
- Codebook. (Höfler et al., 2024)
- Supplementary explanatory materials. (Höfler et al., 2024)

References

Alvarez-Vargas, D., Braithwaite, D., Lortie-Forgues, H., Moore, M., Wan, S., Martin, E., & Bailey, D. H. (2023). Hedges, mottes, and baileys: Causally ambiguous statistical language can increase perceived study quality and policy relevance. *PLoS One, 18*(10), Article e0286403. https://doi.org/10.1371/journal.pone.0286403

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96*, Article 104159. https://doi.org/10.1016/j.jesp.2021.104159

Buchanan, E. M. (2024). *ViSe: Visualizing sensitivity. R package version v0.0.1. Zenodo.* https://doi.org/10.5281/zenodo.10698072

Calin-Jageman, R. J. (2018). The new statistics for neuroscience majors: Thinking in effect sizes. *Journal of Undergraduate Neuroscience Education, 16*(2), E21–E25. https://doi.org/10.31779/471.2.212

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Erlbaum.

D'Agostino McGowan, L. (2022). Sensitivity analyses for unmeasured confounders. *Current Epidemiology Reports, 9*(4), 361–375. https://doi.org/10.1007/s40471-022-00308-6

Daníelsdóttir, H. B., Aspelund, T., Shen, Q., Halldorsdottir, T., Jakobsdóttir, J., Song, H., Lu, D., Kuja-Halkola, R., Larsson, H., Fall, K., Magnusson, P. K. E., Fang, F., Bergstedt, J., & Valdimarsdóttir,

U. A. (2024). Adverse childhood experiences and adult mental health outcomes. *JAMA Psychiatry, 81*(6), 586–594. https://doi.org/10.1001/jamapsychiatry.2024.0039

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science, 5*(3). https://doi.org/10.1177/25152459221106366

Eberhard, K. (2023). The effects of visualization on judgment and decision-making: A systematic literature review. *Management Review Quarterly, 73*(1), 167–214. https://doi.org/10.1007/s11301-021-00235-8

Fox, M. P., MacLehose, R. F., & Lash, T. L. (2021). *Applying quantitative bias analysis to epidemiologic data*. Springer International Publishing. https://doi.org/10.1007/978-3-030-82673-4

Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's d: comparison of two methods. *PloS One, 6*(4), Article e19070. https://doi.org/10.1371/journal.pone.0019070

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models* (Higher Education). Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Glymour, C., Zhang, K., & Spirtes, P. (2019) Review of causal discovery methods based on graphical models. *Frontiers in Genetics, 10*, Article 524. https://doi.org/10.3389/fgene.2019.00524

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 168*(2), 267–306. https://doi.org/10.1111/j.1467-985X.2004.00349.x

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

Guo, G., Karavani, E., Endert, A., & Kwon, B. C. (2023). Causalvis: Visualizations for causal inference. In Proceedings of the 2023 CHI conference on human factors in computing systems (pp. 1–20). https://www.bckwon.com/pdf/causalvis.pdf

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

Hernan, M., & Robins, J. M. (2020). *Causal inference: What if (the book)*. Harvard University Faculty Website: Miguel Hernan. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*(5), 295–300. https://doi.org/10.1177/003591576505800503

Höfler, M., Buchanan, E. M., & Pronizius, E. (2024). *How large must an associational mean difference be to support a causal effect?* [OSF project page containing R code, metadata, codebook, vignettes, and supplementary explanatory materials]. OSF. https://osf.io/s9t8e/

Höfler, M., Scherbaum, S., Kanske, P., McDonald, B., & Miller, R. (2022). Means to valuable exploration: I. The blending of confirmation and exploration and how to resolve it. *Meta-Psychology, 6*. https://doi.org/10.15626/MP.2021.2837

Höfler, M., Trautmann, S., & Kanske, P. (2021). Qualitative approximations to causality: Non-randomizable factors in clinical psychology. *Clinical Psychology in Europe, 3*(2), 1–12. https://doi.org/10.32872/cpe.3873

Kisely, S., Abajobir, A. A., Mills, R., Strathearn, L., Clavarino, A., & Najman, J. M. (2018). Child maltreatment and mental health problems in adulthood: birth cohort study. *British Journal of Psychiatry, 213*(6), 698–703. https://doi.org/10.1192/bjp.2018.207

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(12), 1524–1529. https://doi.org/10.1097/00004583-200312000-00022

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and registered reports. *Evidence-Based Toxicology, 2*(1), Article 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Mair, P., & Wilcox, R. (2020). Robust statistical methods in R using the WRS2 package. *Behavior Research Methods, 52*(2), 464–488. https://doi.org/10.3758/s13428-019-01246-w

Mathur, M. B., & VanderWeele, T. J. (2022). Methods to address confounding and other biases in meta-analyses: Review and recommendations. *Annual Review of Public Health, 43*, 19–35. https://doi.org/10.1146/annurev-publhealth-051920-114020

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

Pastore, M., & Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology, 10*, Article 1089. https://doi.org/10.3389/fpsyg.2019.01089

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st ed.). Basic Books.

Popper, K. (2002). *The logic of scientific discovery*. Routledge.

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

Severini, T. A. (2000). *Likelihood methods in statistics*. Oxford University Press.

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

Torchiano, M. (2020). *effsize: Efficient effect size computation* (0.8.1). https://cran.r-project.org/web/packages/effsize/index.html

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

Methodology
2024, Vol. 20(4), 318–335
https://doi.org/10.5964/meth.14579

PsychOpen GOLD