Behavior Research Methods (2024) 56:5667–5692
https://doi.org/10.3758/s13428-023-02298-9

ORIGINAL MANUSCRIPT

What does a population-level mediation reveal about individual people?

Paul C. Bogdan[1] · **Víctor H. Cervantes**[2] · **Michel Regenwetter**[2,3,4]

Accepted: 16 November 2023 / Published online: 29 December 2023

Abstract

Mediation analysis investigates the covariation of variables in a *population* of interest. In contrast, the resolution level of psychological theory, at its core, aims to reach all the way to the behaviors, mental processes, and relationships of *individual persons*. It would be a logical error to presume that the population-level pattern of behavior revealed by a mediation analysis directly describes all, or even many, individual members of the population. Instead, to reconcile collective covariation with theoretical claims about individual behavior, one needs to look beyond abstract aggregate trends. Taking data quality as a given and a mediation model's estimated parameters as accurate population-level depictions, what can one say about the number of people properly described by the linkages in that mediation analysis? How many individuals are exceptions to that pattern or pathway? How can we bridge the gap between psychological theory and analytic method? We provide a simple framework for understanding how many people actually align with the pattern of relationships revealed by a population-level mediation. Additionally, for those individuals who are exceptions to that pattern, we tabulate how many people mismatch which features of the mediation pattern. Consistent with the person-oriented research paradigm, understanding the distribution of alignment and mismatches goes beyond the realm of traditional variable-level mediation analysis. Yet, such a tabulation is key to designing potential interventions. It provides the basis for predicting how many people stand to either benefit from, or be disadvantaged by, which type of intervention.

Keywords Individual differences · Mediation · Scientific reasoning fallacies · Theoretical scope

Introduction

Unlike political science and sociology, which prioritize aggregate properties of population distributions, the science of psychology aims to reach a level of resolution all the way to individual people. Yet, contemporary strategies for analyzing psychological data, such as with *t* tests or correlations, often operate at the population level. Even when they are rigorously applied to the highest quality data, such analyses can be disconnected from psychological theory of the individual. In this paper, we aim to bridge the model–theory gap for mediation analysis. Mediation testing is a prominent analytic procedure across many areas of psychological research. Mediation analysis aims to shed light on how an independent variable (X) indirectly impacts a dependent variable (Y) via one or more mediator variables (M). The analysis typically involves fitting regressions that connect the variables, measuring the indirect effect as the product of regression coefficients, and performing frequentist tests to assess the statistical significance of the indirect effect (Baron & Kenny, 1986; MacKinnon et al., 2002; Lee et al., 2021; for a non-regression method, see Imai et al., 2010). This procedure provides the statistical underpinnings for many psychological theories. Rather than insist on individual-level research, we connect population-level models to individual-level theory by unpacking what information a population-level mediation model already offers about the individuals who make up that population.

Formally, we treat the mediation model as a jointly distributed family of random variables that captures a population-level joint distribution. Individual behavior is treated as individual realizations of these random variables when drawing a person at random from the population. Hence, we do not consider within-subject (within-individual) mediation. Instead, we are interested in what the population-level distribution tells us about what individual draws from

✉ Paul C. Bogdan
pbogda2@illinois.edu

Extended author information available on the last page of the article

such a distribution will look like. As an analogy, consider the distinction between an optical and a digital zoom. While one cannot apply an optical zoom to a digital picture that has already been taken, a digital zoom can still magnify details that are otherwise not visible. Given a population-level mediation analysis, we aim to zoom in as far as we can to see what it tells us about individual people.

Our approach is mostly conceptual and theoretical. Throughout this paper, we never question the empirical paradigm, the data quality, operationalizations of constructs, the replicability of findings, the methods, sample sizes, sample representativeness, or parameter estimates. We take all of these at face value and focus on the scope of a mediation model as a theory about people: Using published mediation studies as illustrative examples, we ask how many individuals are accurately described by the aggregate relationships an analysis depicts. Among those people who are exceptions to these relationships, we ask how many violate the pattern of trends in what way. This focus diverges from typical conceptualizations of mediation results, which often focus on the proportion of an effect that is mediated by other variables (MacKinnon, 2012).

This paper also is not meant to be a critique of mediation analysis but rather to extract more information than is common in the literature. In the process, we presume that all technical and distributional assumptions underlying mediation analyses are met perfectly.[1,2]

Contrasting ergodicity research, our approach does not conceptualize individuals as having their own distributions. Instead, to derive information about individuals from a population-level distribution, we treat each individual as a single sample point in that population, not as a joint distribution of its own (similar to Bergman & Magnusson, 1997; von Eye & Bergman, 2003; von Eye et al., 2009; von Eye & Wiedermann, 2022; Wiedermann & von Eye, 2021). Because we abstract away from within-person covariation of variables, ergodicity, while important in its own right, is orthogonal to our message.

The most closely related prior work is an empirical study by Grice et al. (2015). The authors re-analyzed an existing dataset, which had shown a significant mediation. In their re-analysis, Grice et al. (2015) tabulated how many individuals followed a predicted pattern of high (above-median) and low (below-median) measurements across the mediation variables. However, we broaden the scope from a single pattern to all possible patterns. Additionally, we shift from a single two-step mediation in a single dataset to a conceptual and theoretical framework for better understanding virtually any mediation model, without requiring access to individual data. We also broaden the scope from a single definition of what constitutes high or low values to a general definition.

To build some basic intuition for our approach, briefly consider the theory that high anxiety goes hand-in-hand with low academic performance. Taking the empirical paradigm and the measures of anxiety and performance at face value, a 'perfect' negative correlation ($\rho = -1$) implies that anyone with high anxiety shows low academic performance, and everybody with low anxiety displays high academic performance. Owens et al. (2012) reported an empirical correlation of $-.15$ between these two constructs. The usual interpretation of such correlations is that 'measuring one variable helps account for variance in the other variable.' What does that mean for psychological theory?

Treating Owens et al.'s reported correlation as a valid and accurate population correlation, what does that correlation tell us about individual people? Assume that, at the population level, anxiety and academic performance are continuous variables with a bivariate normal distribution, and that each individual can be characterized by a single value for each variable jointly drawn from that distribution.

With those assumptions, and focusing on above/below median values for now, a population correlation of $-.15$ means that 27.5% of the population (Kendall & Stuart, 1958) experiences a combination of above median anxiety and below median academic performance. Another 27.5% of people exhibit below median anxiety and above median academic performance. In other words, far from applying to everyone, it is only the case for 55% of individuals that above (respectively below) median anxiety goes hand-in-hand with below (resp. above) median academic performance. Framing correlations as binary contingencies such as these permits easier intuition (similar to the "Common Language" effect size by McGraw & Wong, 1992) and allows us to consider the joint contingencies implied by mediation models. Also, note that while we adopt the language of mediation analysis, we focus on co-occurrences of phenomena only: We are deliberately agnostic about the presence or absence of causal links among these phenomena.

Aiming to dig deeper into the pathway that may link anxiety with academic performance, Owens et al. (2012) carried out a mediation analysis (see our Fig. 1). They reported that increased anxiety predicted decreased working memory and that, in turn, low working memory was associated with low academic performance. Intuitively, adding the construct of working memory to the analysis helps us better understand the relationship between anxiety and academic

[1] While this paper's focus is specific to mediation analysis, it is not meant to advocate for or against mediation. The utility of this technique, such as for drawing conclusions about causality, is left for debate by other papers (Danner et al., 2015; Fiedler et al., 2011, 2018; Tate, 2015).

[2] Our focus and approach notably differ from research on generic disconnects between person-level and population-level distributions. Much of this latter literature has focused on ergodicity, e.g., the relationship between across-subject correlations and within-subject correlations (Fisher et al., 2018; Molenaar & Campbell, 2009), warning that the two can be disconnected from each other.

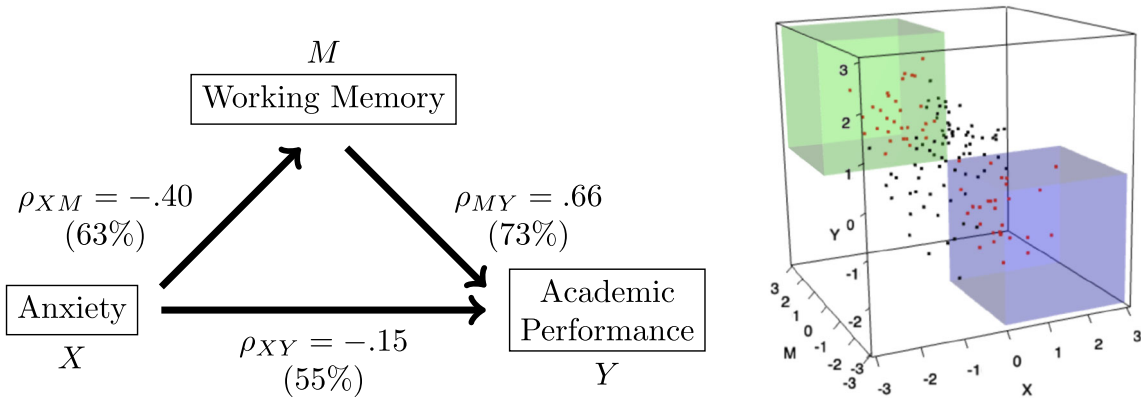

Fig. 1 Working memory as a mediator between anxiety and academic performance, adapted from Owens et al. (2012). Note: We treat Owens et al.'s sample correlations (r) as true population correlations (ρ). All of our analyses rely on multivariate normal distributions and hence use correlation coefficients, contrasting standard mediation-testing procedures, which generally involve regression coefficients. The population percentages on the left were calculated via Eq. 1. Only 45% of individuals in the population (right, red dots) fall into the shaded cubes. See Table 1 for more details. The points in the scatterplot are artificial data simulated from the correlations

performance. However, logically, adding that variable to the analysis cannot increase the number of individuals described accurately by the verbal statement above. That number can only decrease because in addition to explaining how anxiety and performance vary together, the model now also needs to theorize how both vary jointly with memory.

We later show that, taking Owens et al.'s parameter estimates at face value, less than a quarter of the population has the combination of above average anxiety, below average working memory, and below average academic performance (dots in bottom-front cube in the right display of Fig. 1). Combining these with the individuals who have the mirror-image pattern of below average anxiety, above average working memory, and above average academic performance (dots in the top-back in Fig. 1), only 45% of individuals in such a population match the stylized[3] aggregate-level interpretation of the mediation.

As we move from the relationship between two variables to three variables, we move from accurately describing 55% of individuals, already far from everyone, to only 45%, even less than half the population. More generally, adding depth and nuance to a verbal theory by adding mediators in a mediation analysis comes at the cost of reducing the scope of that verbal theory, as fewer and fewer individuals satisfy the combination of stylized phenomena that the mediation model embodies at the aggregate. Recall that we assume, throughout, that the mediation model provides a fully correct representation of the population of interest. Only the stylized theory about anxiety, memory, and performance is limited in scope.

Notably, this example also highlights a difference between our focus and that of ergodicity research on individual- vs. population-level differences (Fisher et al., 2018; Molenaar & Campbell, 2009). Strictly speaking, ergodicity operates at the level of a given single mediation link, i.e., if ergodicity holds for one link, then the population correlation matches a corresponding individual-level correlation. In contrast, we are concerned with how mediation models *combine multiple population-level correlations* and how the combinations decompose into different configurations for individual values. Under this lens, our analyses do not need to measure within-subject covariation to inform inferences about individuals because our focus is specifically on patterns of co-occurrence in individuals.

The remainder of the paper is organized as follows. In the next section, we review scientific reasoning fallacies and how they could affect our interpretation of mediation analyses. In the "Two-step mediation" section, we consider what single mediations, made up of two relationships, $X - M$ and $M - Y$, tell us about individual people. The section titled "Three-step mediation" explores how adding more depth to a mediation, namely creating a sequence of two mediators, affects its theoretical scope by giving individuals many more ways to mismatch the overall pattern of aggregate behavior. We broaden the horizon in the "Generalizations" section by considering more mediation steps, by modeling individuals with particularly high or low values on the variables of interest, and ways to relax multivariate normal distribution

[3] According to `dictionary.cambridge.org/dictionary/english/stylized` the word *stylized* means "represented in a way that simplifies details rather than trying to show naturalness or reality." "If something is stylized, it is represented with an emphasis on a particular style, especially a style in which there are only a few simple details." This is how we use the term throughout the paper.

assumptions. We end with conclusions, open problems, and future directions.

In order to make this paper maximally accessible to a broad readership, the main text relies on illustrative examples, and most of our technical results are relegated to an Appendix. An online shinyapp (https://herulor.shinyapps.io/MediationApp/, available as downloadable R code at https://github.com/herulor/MediationApp) allows the reader both to reproduce our analyses as well as mimic them using two-step or three-step mediation models and associated correlations of their interest. Instructions for the use of the shinyapp are provided in the supplemental materials. We also discuss ideas related to full mediation and cases entailing no mediation in the online supplemental materials.

Scientific reasoning fallacies and how they affect mediation

Logical fallacies can enter into psychological science at many different levels, especially through focus on aggregate data and neglect of differences from one individual to another. The issue of conflating population-level theory and individual behavior matters across a broad variety of research programs. This distinction deserves particular caution for theories composed of multiple parts, as those theories provide fertile ground for leaps of scientific reasoning. For instance, pooled data remain ubiquitous in individual decision research, even though the potential disconnect between individual and collective preference patterns was already noted as early as the 18th century (Condorcet, 1785). Contemporary research has highlighted how certain commonly used behavioral measures will aggregate information so much that the output becomes logically disconnected from the hypothetical constructs of interest, such as individual preferences (Regenwetter & Robinson, 2017, 2019). Decision-making models designed to describe aggregate behavior in a population, in extreme cases, may fail to describe even a single individual (Chen et al., 2020). For a recent exchange about scientific conjunction fallacies in the context of behavioral decision research, see Regenwetter et al. (2022), Kellen (2022), Scheibehenne (2022), Erev & Feigin (2022) and Regenwetter & Robinson (2022).

Estes & Maddox (2005), Kellen & Klauer (2019) and others have pointed out similar problems across a wide range of psychological paradigms. Overreliance on aggregate information can have undesirable consequences. For example, classroom learning interventions may be found to improve average test scores and thus deemed a success, yet they may actually widen achievement gaps if the benefits do not reach underperforming students (Konstantopoulos et al., 2019; Wiedermann et al., 2022).

Some aggregation fallacies can be avoided by considering behavioral patterns at the level of individuals, which is the focus of the *person-oriented research paradigm*. This is an analytic and conceptual approach that frames psychological phenomena in terms of the number of individuals showing a given set of characteristics rather than in terms of variables and population-level statistics (Bergman & Magnusson, 1997; von Eye & Bergman, 2003; von Eye et al., 2009; von Eye & Wiedermann, 2022; Wiedermann & von Eye, 2021; Sterba & Bauer, 2010). For mediation analysis, the person-oriented approach entails investigating the proportion of individuals that match the patterns proposed by a model. For instance, if a model proposes that X enhances Y by upregulating M, then scholars should examine whether a large proportion of individuals display high values of X, M, and Y or low values of X, M, and Y (Collins et al., 1998). Only in such individuals can one conclude that M represents the hypothesized mechanism by which X influences Y. In contrast to the person-oriented approach, traditional *variable-oriented* mediation methods, which employ regressions to link X, M, and Y do not incorporate information about co-occurrence within individuals. Instead, traditional results leave ambiguity, such as about whether the individuals showing high X and high M indeed overlap with those showing high M and high Y.

Person-oriented methods, such as *configural frequency analysis* (CFA) focus on the co-occurrences of outcomes, and, in particular, can be applied to test for the presence of mediation (Bergman & Magnusson, 1997; von Eye & Bergman, 2003; von Eye et al., 2009; von Eye & Wiedermann, 2022; Wiedermann & von Eye, 2021). These analytic strategies focus on discrete variables and involve tabulating the proportion of individuals with each possible configuration of values (such as high X, high M, and high Y), along with evaluating and comparing models that predict the proportion of individuals who adhere to each pattern. This earlier research has demonstrated how one can test statistical significance in the context of person-oriented methods and how one can use person-oriented methods to avoid the types of scientific fallacies that arise when overinterpreting population-level statistics. The present research applies similar principles, tabulating co-occurrences in the presence of interdependencies. However, we consider continuous variables as they arise in mediation. Our focus is not on assessing statistical significance. In fact, we assume that the mediation model (as a multivariate distribution) models all members of the population without exceptions. Rather, by applying person-oriented ideas, we aim to unpack what a traditional population-level mediation result implies

at the level of the individuals who make up that population. We also highlight the importance of considering individuals and co-occurrences when interpreting a mediation result.

Of particular interest for this paper is a line of research that pointed at logical pitfalls in combining findings from multiple studies into a body of evidence. Davis-Stober & Regenwetter (2019) portrayed this idea as a 'paradox' of converging evidence, regarding psychological theories that make multiple predictions: If a researcher generates a novel prediction based on a theory and then identifies experimental evidence of this prediction, this is commonly taken as support for the theory in general. However, additional predictions and studies, even those that yield significant effects, can actually narrow the scope of a theory. With every additional prediction, fewer individuals may satisfy all of the theory's predictions jointly. Davis-Stober & Regenwetter (2019) documented this challenge for studies that rely on Cohen's *d* effect sizes. Mediation generally is built on extracting patterns from combinations of regression weights(Alwin & Hauser, 1975; Baron & Kenny, 1986; MacKinnon et al., 2002; Imai et al., 2010; Lee et al., 2021). This raises similar concerns as combining other effect sizes, like the Cohen's *d* effect sizes we just mentioned, irrespective of whether one combines these weights from within one and the same, or different, sets of data.

Consider a theory that suggests a link between agreeableness and empathy. It may hypothesize that agreeableness is associated with a willingness to help others (Graziano et al., 2007). It may also suggest that agreeableness is associated with heightened physiological responses to others' pain (Courbalay et al., 2015). However, it has also been reported that high personal distress towards others' pain can impair people's helping behavior (Thomas, 2013). Because these studies report population-level trends, the three findings are not at all contradictory. Yet, it is far from clear how many individual people fit that complex pattern of behavior: How many agreeable individuals are very willing to help others while having both a heightened physiological response to others' pain and while dampening responses to others' pain when helping them?

Asking these questions helps protect us against fallacies of sweeping generalization (invalid lines of reasoning from the general to the specific), such as misinterpreting aggregate trends as behavior of many individuals. It also helps us avoid fallacies of composition (invalid lines of reasoning from the specific to the general), such as taking it as a given that prominent phenomena must co-occur. The answers are key to understanding the scope of highly specific and nuanced theories where seemingly conflicting predictions can coexist.

Intuitively, mediation analyses are easy to interpret, as one can readily imagine chains of variables influencing one another. However, that intuition may also misguide users. For instance, there is literature and an ongoing debate about how to avoid correlation-causation fallacies.[4] Various guidelines have become required practices by a number of psychology journals. These requirements are not necessarily meant to dissuade researchers away from mediation analyses. Rather, they aim to ensure a high standard for accepted practice in mediation testing. As we alluded to in the Introduction, our analyses here are outside the realm of these recommendations. We deliberately aim to remain agnostic about mediation as a method of choice for answering any particular scientific question. Instead, we elucidate the theoretical scope of well-fitting mediation models and what such mediations tell us about psychological theory. Specifically, because a mediation pathway is a conjunction of links, we must be on the lookout for conjunction fallacies or similar fallacies of composition, so as to better understand the theory of individual behavior that goes along with a mediation model as a theory of aggregate patterns of behavior[5].

Two-step mediation

In-depth example

We return to our first example from the Introduction to look more closely at two-step mediation. Owens et al. (2012) investigated the effects of emotional processes on academic performance. They concluded, "*Worry and central executive processes mediated the link between negative affect and academic performance*" (Owens et al., 2012, Abstract). The authors specifically found that increased anxiety predicted decreased working memory capabilities ($r = -.40$, $p < .05$), that working memory was linked to academic performance ($r = .66$, $p < .01$), and that the total effect between anxiety and academic performance was not significant ($r = -.15$, $p > .05$). Based on separate analyses using regressions and the Sobel test (Sobel, 1982), the paper reported that these relationships yielded significant evidence of mediation (see Fig. 1).

[4] For instance, scholars are generally advised to consider different causal directions, only use mediation to analyze longitudinal variables, and/or use mathematical approaches designed to specifically investigating causality (Danner et al., 2015; Fiedler et al., 2011, 2018; Maxwell & Cole, 2007; Maxwell et al., 2011; Nguyen et al., 2021; Thoemmes & Lemmer, 2019). For a different line of research that questions the viability of multi-variable causal models in general, on conceptual, logical, policy, and statistical grounds, see Trafimow (2017) and Saylors & Trafimow (2020).

[5] In doing so, it is important to keep in mind that we retain the formalism of mediation at the population level, and we eschew introducing individual-level mediation models. We discuss what one can learn from treating individuals as individual joint realizations of the random variables that make up the mediation model.

For most of this paper, we look at mediation analysis through the lens of multivariate normal distributions, which we construct from papers' reported correlation matrices. Our strategy may appear to diverge from standard mediation procedures, which evaluate regressions rather than correlations. However, note that any regression's standardized weights can be ascertained perfectly based on the correlations among its variables. Here, we start with the assumption that anxiety (X), working memory (M), and academic performance (Y) are jointly normally distributed in the population of interest[6]. Note also that the lack of statistical significance for Owens et al.'s correlation between memory and performance could be due to the absence of an effect, i.e., a zero population correlation $\rho = 0$. Alternatively, there may not have been enough observations to rule out that $r = -.15$ might have originated from $\rho = 0$. Regardless, we take the three sample correlations at face value and consider what it would mean to have population correlations $\rho_{XM} = -.4$ between anxiety and working memory, $\rho_{MY} = .66$ between working memory and academic performance, and $\rho_{XY} = -.15$ between anxiety and academic performance.

We anchor our terminology on the population means as reference points against which we can compare individuals. Denoting an above average value of X, M, or Y as *high* and a below average value as *low*, we show all possible combinations of such values for the three variables in Table 1. Since the three variables X, M, and Y are normal (the mean and median match each other), 50% of people have a high value and the other 50% have a low value on any one of these variables, separately. If we had 'perfect' correlations across the board, i.e., $\rho^*_{XM} = -\rho^*_{MY} = \rho^*_{XY} = -1$, then all individuals with above average anxiety would also have below average working memory and below average academic performance. Analogously, all individuals with below average anxiety would have high working memory and high academic performance. Each of these groups would make up half of the population, and together these two scenarios would capture the entire population. In contrast, as the first two rows of Table 1 show, less than half the population falls into either of those two scenarios. Notably, we do not mean to imply that 50% should be a critical cutoff, but rather we want to point out that even this basic level is not reached, which speaks to how the model should be interpreted.

Table 1 shows all possible matches and mismatches of the links in the mediation path reported by Owens et al. (2012). The table bears resemblance to those reported by studies that use CFA (von Eye & Bergman, 2003; von Eye et al., 2009; von Eye & Wiedermann, 2022; Wiedermann & von Eye, 2021).

[6] Recall that there are no individual-level random variables or distributions.

Table 1 Proportion of the population matching or mismatching the mediation pattern $X \uparrow M \downarrow Y \downarrow$ of Owens et al. (2012)

	Mediation Path Links Violated	$X \uparrow$ Anxiety	$M \downarrow$ Working Memory	$Y \downarrow$ Academic Performance	Proportion of the Population	
A1	No violation	high	low	low	.227	.454
A2		low	high	high	.227	
B1	XM	high	high	high	.138	.275
B2		low	low	low	.138	
C1	MY	high	low	high	.088	.177
C2		low	high	low	.088	
D1	XM; MY	high	high	low	.047	.094
D2		low	low	high	.047	

Note. We treat the sample correlations of Owens et al. (2012) as accurate population correlations ($\rho_{XM} = -.4$, $\rho_{MY} = .66$, and $\rho_{XY} = -.15$). Above average (i.e., above median) is labeled as "high," below average (i.e., below median) is labeled as "low." Comparing to the right side of Fig. 1, subpopulation A1 corresponds to the points in shaded cube on the lower front in (X, M, Y) space, whereas subpopulation A2 corresponds to the dots in the shaded cube in the upper back. A version of this table that provides marginal confidence intervals based on treating the results of Owens et al. (2012) as sample correlations, is provided in the online supplemental materials as Table B1

We write $X \uparrow M \downarrow$ to denote that the mediation declares high (resp. low) anxiety to go hand-in-hand with low (resp. high) working memory.[7] Likewise, $M \uparrow Y \uparrow$ denotes the mediation finding that high (resp. low) working memory is linked with high (resp. low) academic performance. Each line in Table 1 shows a pattern of high (resp. low) values for the three variables. In each line, it shows whether and how the pattern violates the mediation pattern $X \uparrow M \downarrow Y \downarrow$. Each line also shows what proportion of the population satisfies that pattern, according to the trivariate normal distribution underlying the mediation model. Combining the first two lines in the table, 45% of individuals, less than half the population, display the $X \uparrow M \downarrow Y \downarrow$ pattern conveyed by the mediation. These individuals either have high X, together with low M and low Y; or they have low X together with high M and high Y.

Everyone else is an exception to that pattern. For instance, pattern D1 denotes the individuals with high anxiety yet also high working memory, combined with low academic performance. This group makes up almost 5% of the population. While they are the smallest constituent groups in the table, individuals in D1 (high X, high M, and low Y) and individuals in D2 (low X, low M, and high Y), together, nonetheless

[7] Notice that the relationship $X \uparrow M \downarrow$ is the same as the relationship $X \downarrow M \uparrow$. Remember, also, that we do *not* consider within-person variation.

make up 9.4% of the population. This means that nearly a tenth of the population *completely mismatches* the mediation pattern $X \uparrow \ M \downarrow \ Y \downarrow$ in that they mismatch both the XM and the MY relationship patterns.

To see how Table 1 and the left side of Fig. 1 are related, notice that the proportion of people with patterns A1, A2, C1, or C2 ($.227 + .227 + .088 + .088 = .63$) aligns with our earlier observation that $\rho_{XM} = -.40$ translates into 63% of the population (see Fig. 1) satisfying the 'negative' relationship $X \uparrow \ M \downarrow$ between anxiety and working memory. Similarly, the proportion of people in the population who are in A1, A2, B1, or B2 ($.227 + .227 + .138 + .138 = .73$) aligns with the finding that $\rho_{MY} = .66$ translates into 73% of the population satisfying the 'positive' relationship $M \uparrow \ Y \uparrow$ between working memory and academic performance. To see how Table 1 and the right side of Fig. 1 are related, note that 45% of the dots lie in the shaded cubes in that figure. In all, while the pattern $X \uparrow \ M \downarrow \ Y \downarrow$, that emerges from the mediation model, indeed delineates the relatively largest fraction of the population (45% of all individuals), the overall mediated effects primarily reflect properties of the population and not so much of individuals themselves.

Consequently, interventions that aim to help people overcome adversity may benefit far fewer individuals than one might anticipate from a cursory interpretation of the mediation analysis. For example, interventions that aim to improve the academic performance of students who are both very anxious and have low working memory, should take into account that less than a quarter of all students are in row A1 of the table, matching the aggregate pattern of undesirable traits. They should also take into account that the roughly 9% of students who are in row C1, enjoy high academic performance despite having high anxiety and low working memory. If administered to entire classrooms, this intervention potentially only reaches about a quarter of all students (row A1) and might be somewhat of a mismatch for another 9% (row C1). On the other hand, an intervention that aims to improve academic performance by alleviating working memory deficits related to anxiety would be much more efficient if targeted at students who show high anxiety and poor academic performance. Among them, 82.8% (row A1 divided by the sum of A1 and D1) of students have impaired working memory and thus stand to benefit. While some interventions may already be targeted as a matter of procedure, the table helps to quantify how many people stand to benefit which way.

Tabulating the mediation predictions, like in Table 1, could be a key step towards avoiding misinterpretations of the results and could protect many individuals against unintended consequences in interventions. As we will show in other examples, the subpopulation that matches the mediation pattern tends to shrink even further as the mediation links weaken or as we insert more links.[8]

As an aside, it should be noted that, in psychological research, a sample correlation between measurements will inherently underestimate the population relationship between their underlying constructs because the measurements are not perfectly reliable (i.e., there is measurement error). For example, if the sample correlation between X and Y is .3, but X and Y each have reliability of .75, then the best estimate of the population correlation between the underlying constructs is .4 (see e.g., Thorndike et al., 2010). Analogous to how worsening data quality will weaken correlations (i.e., pull correlations toward $\rho = 0$), this issue will cause differences in pattern probabilities to shrink (e.g., for a two-step mediation, pull all eight table rows towards .125). To account for this, readers may adjust their correlations upward as they see fit to account for reliability issues. Our results do not include any such adjustments because we treat the population correlations as known. As we show further below, such adjustments typically only affect the core qualitative conclusions if all of the correlations approach ± 1.0.

General results

We are now ready to state general results for two-step mediations. Suppose that (X, Y, M) has a trivariate normal distribution. Note again that the median in a normal distribution is the same value as the mean. We employ a notation that gives us the flexibility to cover different patterns. Starting with X, we use ✓ to denote a value above the median and ⊗ to denote a value below the median for X. Moving along the mediation path (here, from X to M to Y), depending on the sign of the correlation connecting adjacent variables, ✓ can denote a value either above or below the median for variables other than X, with ⊗ denoting a value on the opposite side of the median: For two adjacent variables in a mediation pathway that are positively correlated, we will set ✓ for these variables in such a way that they denote the same side of the median. This can be written as follows, with $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ as the corresponding dichotomized variables taking the values

[8] The latter trend is similar to a trend, discussed by Trafimow (2017); Saylors & Trafimow (2020), that, with every additional variable added in a causal model, the number of correlations grows ever more quickly, and with it, the number of inferences about causal relationships. As Saylors et al. explain, this means that the confidence a scholar can have in inferring a 'correct' causal model from data will deteriorate rapidly with increasing numbers of variables.

✓ and ⊗,

$$
\begin{aligned}
\widetilde{X} = \checkmark &\iff X \geq x_{50},\\
\widetilde{X} = \otimes &\iff X < x_{50},\\
\widetilde{M} = \checkmark &\iff \text{sign}(\rho_{XM})M \geq m_{50},\\
\widetilde{M} = \otimes &\iff \text{sign}(\rho_{XM})M < m_{50},\\
\widetilde{Y} = \checkmark &\iff \text{sign}(\rho_{XM}\rho_{MY})Y \geq y_{50},\\
\widetilde{Y} = \otimes &\iff \text{sign}(\rho_{XM}\rho_{MY})Y < y_{50}.
\end{aligned}
$$

For two adjacent variables in a mediation pathway that are negatively correlated, we set ✓ for those variables in such a way that they denote opposite sides of the median. For the Owens et al. mediation, the pattern ✓✓ on XM indicates above average anxiety and below average working memory. Then, for MY, the pattern ✓✓ denotes below average working memory and below average performance. The pairwise joint distributions of the dichotomizations are given in the contingency tables in Fig. 2.

This way of labeling allows us to capture many different situations with a single mnemonic notation. For instance, in the mediation of Fig. 1, the pattern of high X combined with low M and with low Y is represented as ✓✓✓. In that example, ⊗ ⊗ ⊗ denotes the mirror-image pattern of low X with high M and high Y. Note that, in a situation where $\rho_{XM} > 0$, $\rho_{MY} > 0$, but $\rho_{XY} < 0$, which is conceptually strange but mathematically possible and has been observed in some empirical data (MacKinnon et al., 2000), the pattern ✓✓✓ denotes above median values on all three variables.

	$\widetilde{M} = \checkmark$	$\widetilde{M} = \otimes$	
$\widetilde{X} = \checkmark$	P_{XM}	$1/2 - P_{XM}$	$1/2$
$\widetilde{X} = \otimes$	$1/2 - P_{XM}$	P_{XM}	$1/2$
	$1/2$	$1/2$	
	$\widetilde{Y} = \checkmark$	$\widetilde{Y} = \otimes$	
$\widetilde{M} = \checkmark$	P_{MY}	$1/2 - P_{MY}$	$1/2$
$\widetilde{M} = \otimes$	$1/2 - P_{MY}$	P_{MY}	$1/2$
	$1/2$	$1/2$	
	$\widetilde{Y} = \checkmark$	$\widetilde{Y} = \otimes$	
$\widetilde{X} = \checkmark$	P_{XY}	$1/2 - P_{XY}$	$1/2$
$\widetilde{X} = \otimes$	$1/2 - P_{XY}$	P_{XY}	$1/2$
	$1/2$	$1/2$	

Fig. 2 Pairwise joint distributions of variables $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$

We start by looking at pairs of variables and their bivariate distributions. Let P_{XM}, P_{MY}, and P_{XY} denote the respective probability of jointly drawing ✓✓ for X and M, drawing ✓✓ for M and Y, and drawing ✓✓ for X and Y, respectively, in random samples drawn from the population.

We can directly compute these probabilities from the bivariate normal distribution (Kendall & Stuart, 1958, p. 351) via

$$
\begin{aligned}
P_{XM} &= \frac{1}{4} + \frac{\arcsin\left(|\rho_{XM}|\right)}{2\pi},\\
P_{MY} &= \frac{1}{4} + \frac{\arcsin\left(|\rho_{MY}|\right)}{2\pi},\\
P_{XY} &= \frac{1}{4} + \frac{\arcsin\left(\text{sign}(\rho_{XM}\rho_{MY})\rho_{XY}\right)}{2\pi}.
\end{aligned}
\tag{1}
$$

The absolute value in the first two equations, and the sign of the product of the correlations in the path, in the third equation, ensure that the probabilities correctly reflect the encoding of ✓ and ⊗ for each variable as described above. For instance, using the correlations that we adopted from Owens et al. (2012), we obtain $P_{XM} = .316$, $P_{MY} = .365$, and $P_{XY} = .274$. Specifically, $P_{XM} = .316$ means that the probability that a randomly drawn person has above average anxiety and below average working memory is .316. Hence, 31.6% of the population has above average anxiety and below average working memory, which also corresponds to ✓✓ for the combination of X and M. Thanks to the symmetries of normal distributions, the probability of drawing a ⊗⊗ pattern is just the same as the probability of drawing a ✓✓ pattern. As a consequence, the probability of drawing either ✓✓ or ⊗⊗ for X and M is $2P_{XM} = .631$; the probability of drawing either ✓✓ or ⊗⊗ for M and Y is $2P_{MY} = .730$, and the probability of drawing either ✓✓ or ⊗⊗ for X and Y is $2P_{XY} = .548$. These proportions correspond to the percentages we give in parentheses in Fig. 1 (left).

Next, taking into account that (X, M, Y) is trivariate normal, we construct Table 2 to represent possible patterns across X, M, and Y.

Table 2 Distribution of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$

	$\widetilde{X}$	$\widetilde{M}$	$\widetilde{Y}$	Joint Outcome Probability
$A1$	✓	✓	✓	$\Pr_{A1}$
$A2$	⊗	⊗	⊗	$\Pr_{A2}$
$B1$	✓	⊗	⊗	$\Pr_{B1}$
$B2$	⊗	✓	✓	$\Pr_{B2}$
$C1$	✓	✓	⊗	$\Pr_{C1}$
$C2$	⊗	⊗	✓	$\Pr_{C2}$
$D1$	✓	⊗	✓	$\Pr_{D1}$
$D2$	⊗	✓	⊗	$\Pr_{D2}$

Note that the sum of $\Pr_{A1}$ and $\Pr_{A2}$ represents the proportion of individuals 'matching' the hypothesized mediation pathway. The eight probabilities in Table 2 are real numbers between 0 and 1. They must satisfy the following seven constraints:

$$\Pr_{A1}+\Pr_{A2}+\Pr_{B1} + \Pr_{B2}+\Pr_{C1}+\Pr_{C2}+\Pr_{D1}+\Pr_{D2}=1 \quad (2)$$

$$\Pr_{A1}+\Pr_{B1}+\Pr_{C1}+\Pr_{D1}=\Pr(\widetilde{X}=\checkmark)=1/2, \quad (3)$$

$$\Pr_{A1}+\Pr_{B2}+\Pr_{C1}+\Pr_{D2}=\Pr(\widetilde{M}=\checkmark)=1/2, \quad (4)$$

$$\Pr_{A1}+\Pr_{B2}+\Pr_{C2}+\Pr_{D1}=\Pr(\widetilde{Y}=\checkmark)=1/2, \quad (5)$$

$$\Pr_{A1}+\Pr_{C1}=\Pr_{A2}+\Pr_{C2}=P_{XM}, \quad (6)$$

$$\Pr_{A1}+\Pr_{B2}=\Pr_{A2}+\Pr_{B1}=P_{MY}, \quad (7)$$

$$\Pr_{A1}+\Pr_{D1}=\Pr_{A2}+\Pr_{D2}=P_{XY}. \quad (8)$$

Equations 2-8 constitute seven linearly independent constraints that leave a single degree of freedom in determining the values of the eight probabilities in the last table. If we write a for the first probability, $\Pr_{A1}$, then the remaining probabilities, $\Pr_{A2},\ldots,\Pr_{D2}$, are completely determined by the probabilities P_{XM}, P_{MY}, and P_{XY}. For example, from $P_{XM}=\Pr_{A1}+\Pr_{C1}$ we obtain that $\Pr_{C1}=P_{XM}-a$. In all, this gives the distribution in Table 3.

By summing probabilities $\Pr_{A1}$ and $\Pr_{A2}$, we see that, given the values P_{XM}, P_{MY}, and P_{XY}, the proportion of people who match the mediation pattern equals

$$P_{XM}+P_{MY}+P_{XY}-1/2. \quad (9)$$

This result applies to all possible joint distributions of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ that can be obtained by uniformly splitting X, M, and Y at their medians.

Moreover, if we assume that the proportions for two mutually opposite patterns of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ are equal[9], for instance if $\Pr_{A1}=\Pr_{A2}$, then it is immediate that $\Pr_{A1}=\frac{1}{2}(P_{XM}+P_{MY}+P_{XY}-1/2)$. We can completely describe the distribution by the probabilities in Table 4. In particular, the probabilities in Table 4 hold if X, M, and Y have a trivariate normal distribution. The online shinyapp (https://herulor.shinyapps.io/MediationApp/) computes the probabilities for patterns obtained by splitting at the medians of the variables X, M, and Y, under two-step mediation, using the expressions in Table 4.

Central to the question of mediation, the probability of drawing an individual with pattern A1, in a uniform random sample from the population, is given by the formula

[9] It suffices to assume the equality for only one of the mirror-image pairs of proportions. The other equalities follow from that.

Table 3 Probability values for the distribution of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$

	$\widetilde{X}$	$\widetilde{M}$	$\widetilde{Y}$	Joint Outcome Probability
$A1$	✓	✓	✓	$\Pr_{A1}=a$
$A2$	⊗	⊗	⊗	$\Pr_{A2}=P_{XM}+P_{MY}+P_{XY}-a-1/2$
$B1$	✓	⊗	⊗	$\Pr_{B1}=1/2-P_{XM}-P_{XY}+a$
$B2$	⊗	✓	✓	$\Pr_{B2}=P_{MY}-a$
$C1$	✓	✓	⊗	$\Pr_{C1}=P_{XM}-a$
$C2$	⊗	⊗	✓	$\Pr_{C2}=1/2-P_{MY}-P_{XY}+a$
$D1$	✓	⊗	✓	$\Pr_{D1}=P_{XY}-a$
$D2$	⊗	✓	⊗	$\Pr_{D2}=1/2-P_{XM}-P_{MY}+a$

$\frac{1}{2}(P_{XM}+P_{MY}+P_{XY}-1/2)$, shown in the last column of Table 4. For the Owens et al. (2012) mediation analysis results, this yields $\frac{1}{2}(.316+.365+.274-1/2)=.227$. Here, 22.7% of individuals have high anxiety, low working memory, and low academic performance; a separate 22.7% have the mirror-image combination of low anxiety, high working memory, and high academic performance. We obtained the other values in Table 1 in similar ways.

The results so far have not yet fully leveraged the fact that we are considering mediation analyses. In particular, a *full mediation* applies when, after taking into account M, the variables X and Y are independent. We can provide similar analyses and tabulations through the lens of full mediation. We report our findings in the online supplemental materials. The results show that full mediation does not necessarily lead to any higher proportions of individuals matching a stylized aggregate theory.

Three-step mediation

To demonstrate how each additional step within a mediation further delineates the scope of a stylized theory, we next examine the three-step developmental psychology mediation

Table 4 Probability values for the distribution of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ with symmetry assumption

	$\widetilde{X}$	$\widetilde{M}$	$\widetilde{Y}$	Joint Outcome Probability
$A1$	✓	✓	✓	$\frac{1}{2}(P_{XM}+P_{MY}+P_{XY}-1/2)$
$A2$	⊗	⊗	⊗	$\frac{1}{2}(P_{XM}+P_{MY}+P_{XY}-1/2)$
$B1$	✓	⊗	⊗	$\frac{1}{2}(-P_{XM}+P_{MY}-P_{XY}+1/2)$
$B2$	⊗	✓	✓	$\frac{1}{2}(-P_{XM}+P_{MY}-P_{XY}+1/2)$
$C1$	✓	✓	⊗	$\frac{1}{2}(P_{XM}-P_{MY}-P_{XY}+1/2)$
$C2$	⊗	⊗	✓	$\frac{1}{2}(P_{XM}-P_{MY}-P_{XY}+1/2)$
$D1$	✓	⊗	✓	$\frac{1}{2}(-P_{XM}-P_{MY}+P_{XY}+1/2)$
$D2$	⊗	✓	⊗	$\frac{1}{2}(-P_{XM}-P_{MY}+P_{XY}+1/2)$

reported by Simpson et al. (2007), who concluded that "*targets classified as securely attached at 12 months old were rated as more socially competent during early elementary school by their teachers. Targets' social competence, in turn, forecasted their having more secure relationships with close friends at age 16, which in turn predicted more positive daily emotional experiences in their adult romantic relationships*" (p. 355; illustrated here in Fig. 3).

Consider the following embedded[10] two-step mediations. First, infant attachment predicted early peer competence ($r = .29$, $p < .01$), which, in turn, predicted success in adult relationships ($r = .27$, $p < .05$). Second, the association between peer competence and adult relationship success was itself mediated by friendship security at age 16. Peer competence predicted friendship security ($r = .37$, $p < .01$), which, in turn, predicted success in adult relationships ($r = .48$, $p < .001$). Together, the four variables yielded a three-step mediation that fit the data well, per the authors' structural equation modeling.

Like before, we take the analyses at face value and adopt the sample correlations as accurate population correlations that capture the covariation across the entire population of interest. Table 5 shows the corresponding population correlations in the top and the associated population proportions in the bottom. For instance, we presume that the sample correlation $r = .29$ between infant attachment (our X variable) and peer competence (our M_1 variable) accurately reveals a population correlation of $\rho_{XM_1} = .29$. For X and M_1, following the conventions introduced earlier, we write ✓✓ to denote high X and high M_1. Then, by Eq. 1, an individual who is randomly drawn from a bivariate normal with $\rho_{XM_1} = .29$ displays ✓✓ with probability .297 and ⊗⊗ with probability .297. In other words, a randomly sampled individual has probability .594 to satisfy the mediation pattern $X \uparrow M_1 \uparrow$ that relates X with M_1. The same logic applies to the rest of the table.

Combinatoric explosion of path links and patterns

Before we look more closely at the numbers, notice how, as we moved from two-step to three-step mediation, we also moved from three correlation coefficients to six. Similarly, as we distinguish between high and low values on each of four variables, we now encounter double the patterns. There are now 16 different binary patterns, of which again just two, now ✓✓✓✓ and ⊗⊗⊗⊗, match the aggregate mediation pattern. This makes it salient that the additional depth of the three-step mediation introduces many more opportunities for individuals to mismatch the stylized pattern of behavior embodied at the aggregate level of the mediation model.

If, as is common in developmental psychology, we view Simpson et al.'s mediation model as a pathway that leads from infancy to adulthood by linking X to M_1, then M_1 to M_2, then M_2 to Y, we can naturally ask what combination of these links correctly describe the developmental path taken by a given individual. Following again the convention to denote a value above the median on X as ✓, a person with high infant attachment, high peer competence, high friendship security, and high relationship success, who has pattern ✓✓✓✓, satisfies every hypothesized relationship of the mediation. On the other hand, a person with high infant attachment, high peer competence, low friendship security and high relationship success, who has pattern ✓✓⊗✓, has navigated this path in a fashion that mismatches two stylized mediation links. They mismatch the link from M_1 to M_2 because they moved from above-median peer competence to below-median friendship security, which constitutes a mismatch with the positive correlation $\rho_{M_1M_2} = .37$ between those two variables. They are also an exception to the stylized link from M_2 to Y because they moved from below-median friendship security to above-median relationship success, which constitutes a mismatch with the positive correlation $\rho_{M_2Y} = .48$. Table 6 shows the patterns in the center and the violated mediation path links on the left. As we have just seen, the pattern ✓✓⊗✓ violates the $M_1 \uparrow M_2 \uparrow$ link as well as the $M_2 \uparrow Y \uparrow$ link in the path from X through M_1 then M_2 to Y. (The table leaves out the arrows to accommodate all possible correlation signs in the general case.) However, mismatching those two links goes hand-in-hand with mismatching a larger number of stylized patterns, implied by single correlations. Using the same example, a person with high infant attachment, high peer competence, low friendship security and high relationship success also mismatches the $X \uparrow M_2 \uparrow$ pattern of Simpson et al. (2007). Furthermore, this person mismatches the aggregate patterns of two embedded two-step mediations, namely $X \uparrow M_2 \uparrow Y \uparrow$ and $M_1 \uparrow M_2 \uparrow Y \uparrow$. Accordingly, the right column of Table 6 shows that a person with the pattern ✓✓⊗✓ violates the XM_2, M_1M_2 and M_2Y patterns, as well as the XM_2Y and M_1M_2Y patterns.

It is noticeable from Table 5 that none of the pairwise patterns, ✓✓ or ⊗⊗, capture large segments of the population. This raises two salient questions: 1) How are the two-step mediations among X, Y, and each mediator affected by the smaller correlations relative to those in the earlier Owens et al. example? 2) How many individual people have outcomes that match the overall pattern $X \uparrow M_1 \uparrow M_2 \uparrow Y \uparrow$ that Simpson et al. (2007) highlight in their findings? Put simply, what is the scope of Simpson et al.'s stylized theory?

[10] We use the term *embedded* to refer to marginals of a full model, not to discuss counterfactuals of what would have happened if a study had omitted one or more variables.

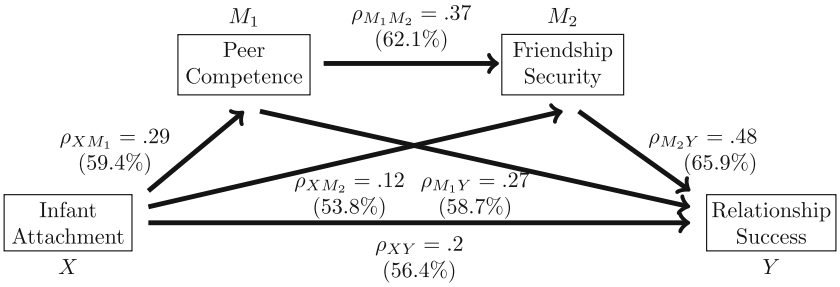

Fig. 3 Three-step mediation in which peer competence (in grades 1–3) and friendship security (at age 16) mediate the relationship between infant attachment (at 12 months of age) and relationship success (at ages 20–23), based on Simpson et al. (2007). *Note.* Percentages indicate the proportion of the population whose values on two variables are on the same side of the median (i.e., both high or both low)

Two-step mediation embedded in the three-step mediation

We start with the first question by considering the marginal trivariate normal distribution on three of the four variables. Table 7 is the direct analogue to Table 1 for $X \uparrow M_1 \uparrow Y \uparrow$, a two-step mediation that looks at how peer competence mediates the relationship between infant attachment and relationship success. Summing the first two rows in the table yields only 37.4%. Barely more than a third of the population abides by the pattern $X \uparrow M_1 \uparrow Y_\uparrow$ suggested at the aggregate level by this two-step mediation. Looking at the bottom two rows of Table 7 reveals that nearly 20% of individuals behave in maximal disagreement with the aggregate pattern of this two-step mediation. As we compare this mediation with that in Fig. 1 and Table 1, we see that reducing $|\rho_{XM}| = .4$ to $|\rho_{XM_1}| = .29$ and $|\rho_{MY}| = .66$ to $|\rho_{M_1Y}| = .27$ (while $|\rho_{XY}| = .15$ changes slightly to $|\rho_{XY}| = .21$) has a notable impact on the pattern probabilities. Not only do far fewer individuals satisfy either ✓✓✓ or ⊗⊗⊗. The most extreme exceptions (in rows D1 and D2) have more than doubled in rate.

Table 5 Pairwise correlations, based on Simpson et al. (2007), among X, M_1, M_2, and Y, and corresponding proportions of the population who satisfy pairwise patterns ✓✓ or ⊗⊗

	Presumed Population Correlations		
	M_1	M_2	Y
X	$\boldsymbol{\rho_{XM_1} = .29}$	$\rho_{XM_2} = .12$	$\boldsymbol{\rho_{XY} = .2}$
M_1		$\rho_{M_1M_2} = .37$	$\boldsymbol{\rho_{M_1Y} = .27}$
M_2			$\rho_{M_2Y} = .48$
	Percent of Population with ✓✓ or ⊗⊗		
	M_1	M_2	Y
X	**59.4%**	53.8%	**56.4%**
M_1		62.1%	**58.7%**
M_2			65.9%

Note. The pairwise correlations among X, M_1, M_2, and Y are given in the top. The proportions of the population who satisfy the corresponding patterns ✓✓ or ⊗⊗ are reported in the bottom. The boldfaced correlations are the basis for the two-step mediation of Table 7. See the text for more specifics

Full three-step mediation

We answer the second question, on how many individuals align with the overall three-step pattern, in the last two columns in Table 8. As we move from the first two lines in Table 7 to the first two lines of Table 8, the proportion of the population that aligns with the aggregate mediation pattern falls further from .372 to only .275. This documents our earlier claim that adding an additional variable to enhance the depth of an analysis cannot increase the number of individuals whose behavior we successfully model with just a single aggregate pattern. Indeed, when we combine all four variables in the three-step mediation, and we need to account for six correlation coefficients, the scope of the mediation pattern $X \uparrow M_1 \uparrow M_2 \uparrow Y \uparrow$, which consists of the two scenarios ✓✓✓✓ and ⊗⊗⊗⊗, encompasses barely more than a quarter of all individual people. 'Perfect' correlations would mean that half the population has above average infant attachment, above average peer competence, above average friendship security, combined with above average relationship success ✓✓✓✓ (and half has the opposite pattern). However, if we take Simpson et al.'s correlations at face value, that pattern describes fewer than 14% of individuals.

Now, consider people with unsuccessful adult relations who, as infants, showed poor attachment. These align with the pairwise pattern implied by $\rho_{XY} = .2$ and form 28.2% of the population (see also Rows $A1$, $C1$, $D1$, $G2$ in Table 7). Even among this subset that shows undesirable outcomes in both X and Y, only a minority (48.9%) also showed bad outcomes on M_1 and M_2. To see that, one needs to divide the 13.8% in row A by the 28.2% who match the $X \uparrow \; Y \uparrow$ pattern (Fig. 3). However, of that same group, almost everyone, namely 84.8%, showed poor social functioning in at least childhood or adolescence: To calculate this percentage, divide the sum across Rows A1, C1, and D1 (i.e., 23.9%) by 28.2%. These sorts of calculations show that few people

Table 6 Relationship between violated mediation path links, value patterns, and mediation pattern mismatch

Mediation Path Links Violated	X	M_1	M_2	Y		X	M_1	M_2	Y	Mediation Pattern Violations
No violation	✓	✓	✓	✓	or	⊗	⊗	⊗	⊗	No violation
XM_1	✓	⊗	⊗	⊗	or	⊗	✓	✓	✓	XM_1; XM_2; XY; XM_1Y; XM_2Y
XM_1; M_1M_2	⊗	✓	⊗	⊗	or	✓	⊗	✓	✓	XM_1; M_1M_2; M_1Y XM_1Y; M_1M_2Y
M_1M_2; M_2Y	⊗	⊗	✓	⊗	or	✓	✓	⊗	✓	XM_2; M_1M_2; M_2Y XM_2Y; M_1M_2Y
M_2Y	⊗	⊗	⊗	✓	or	✓	✓	✓	⊗	XY; M_1Y; M_2Y XM_1Y; XM_2Y
M_1M_2	✓	✓	⊗	⊗	or	⊗	⊗	✓	✓	XM_2; XY; M_1M_2; M_1Y XM_1M_2; XM_1Y; XM_2Y; M_1M_2Y
XM_1; M_2Y	⊗	✓	✓	⊗	or	✓	⊗	⊗	✓	XM_1; XM_2; M_1Y; M_2Y XM_1M_2; XM_1Y; XM_2Y; M_1M_2Y
XM_1; M_1M_2; M_2Y	✓	⊗	✓	⊗	or	⊗	✓	⊗	✓	XM_1; XY; M_1M_2; M_2Y XM_1M_2; XM_1Y; XM_2Y; M_1M_2Y

follow one single trajectory. They provide nuanced insights into the various kinds of developmental trajectories and their prominence. The calculations in Tables 7 and 8 can be reproduced online using the shinyapp. The latter permits the user to input any (mutually compatible) population correlations of their choice to generate similar analyses.

Examining Table 8, as well as our earlier findings, highlights several important insights: 1) Unless all correlations are near ±1, high-level descriptions of mediation findings based on population-level relationships among variables suggest an overly stylized and misleading picture of individual behavior. This second point also means that adjusting the population correlations upwards from the reported sample correlations in the papers we revisit, to accommodate unreliable measurements, can only adjust, but not fundamentally

Table 7 Proportion of the population satisfying or mismatching the component two-step mediation pattern $X \uparrow M_1 \uparrow Y \uparrow$ of Simpson et al. (2007) when treating their sample correlations as accurate population correlations

	Mediation Path Links Violated	X	M_1	Y	Proportion of the Population	
A1	No violation	✓	✓	✓	.186	.372
A2		⊗	⊗	⊗	.186	
B1	XM_1	✓	⊗	⊗	.107	.215
B2		⊗	✓	✓	.107	
C1	M_1Y	✓	✓	⊗	.11	.22
C2		⊗	⊗	✓	.11	
D1	XM_1; M_1Y	✓	⊗	✓	.096	.192
D2		⊗	✓	⊗	.096	

Note. For each variable, ✓ denotes a value above the median; ⊗ denotes a value below the median

Table 8 Proportion of the population satisfying or mismatching the three-step mediation pattern $X \uparrow M_1 \uparrow M_2 \uparrow Y \uparrow$ of Simpson et al. (2007) when treating their sample correlations as accurate population correlations

	Mediation Path Links Violated	X	M_1	M_2	Y	Proportion of the Population	
A1	No violation	✓	✓	✓	✓	.138	.276
A2		⊗	⊗	⊗	⊗	.138	
B1	XM_1	✓	⊗	⊗	⊗	.079	.158
B2		⊗	✓	✓	✓	.079	
C1	XM_1; M_1M_2	⊗	✓	⊗	⊗	.053	.106
C2		✓	⊗	✓	✓	.053	
D1	M_1M_2; M_2Y	⊗	⊗	✓	⊗	.048	.096
D2		✓	✓	⊗	✓	.048	
E1	M_2Y	⊗	⊗	⊗	✓	.05	.10
E2		✓	✓	✓	⊗	.05	
F1	M_1M_2	✓	✓	⊗	⊗	.06	.12
F2		⊗	⊗	✓	✓	.06	
G1	XM_1; M_2Y	⊗	✓	✓	⊗	.043	.086
G2		✓	⊗	⊗	✓	.043	
H1	XM_1; M_1M_2; M_2Y	✓	⊗	✓	⊗	.029	.058
H2		⊗	✓	⊗	✓	.029	

Note. For each variable, ✓ denotes a value above the median; ⊗ denotes a value below the median

change, the resulting picture we draw here. 2) Weaker correlations amplify these problems. 3) As we shift to mediations containing greater numbers of mediating variables, e.g., from two-step to three-step, all else being equal, the number of individuals who match the mediation pattern can only go down. 4) As we detail in the online supplemental materials, full mediation is not a remedy to any of these problems, and cases of full mediation may still show limited proportions of individuals who match aggregate patterns. 5) Analyses like the ones in the above examples provide novel nuance in understanding how mediation, as a model of aggregate covariation of variables, translates into a distribution of heterogeneous individuals who vary in whether they match or how they mismatch that aggregate pattern. 6) Similarly to the last point, these analyses allow mediation users to move beyond the stylized theory that could emerge from a cursory interpretation of aggregate relationships.

Generalizations

Beyond three-step mediations

In this subsection, we demonstrate that expanding a series of mediators will quickly and substantially shrink the subpopulation of people who match the stylized mediation pattern. Elaborate mediations containing four or more steps come at a cost: Put simply, adding depth and nuance to a mediation-based theory narrows that verbal theory's scope, in the sense that fewer and fewer individuals act in accordance with the stylized pattern of relationships and pathways embodied by that verbal theory. For this subsection, without any loss of generality, we consider only positive correlations, which means that ✓ always denotes a value above the median (i.e., above the mean). Hence, like before, for any given single variable, half of all individuals satisfy ✓ and half satisfy the mirror-image ⊗. We consider n many mediators $M_1, M_2, \ldots M_n$ that form a sequence.

The resulting proportions of the population who satisfy ✓✓ . . . ✓ or ⊗ ⊗ · · · ⊗ are provided, for simulated data, in Table 9. For simplicity, we suppose that all correlations are equal, i.e., $0 < \rho = \rho_{XY} = \rho_{XM_i} = \rho_{M_iM_j} = \rho_{M_jY}$ for $1 \leq i \neq j \leq n$. For example, with $n = 1$ (two-step mediation) setting $\rho = .5$ we find that 50%, and setting $\rho = .7$, we find that 62% of the population satisfy ✓✓✓ or ⊗ ⊗ ⊗. This is similar to Table 1, which was based on $|\rho_{XM}| = .4$, $|\rho_{MY}| = .66$, and $|\rho_{XY}| = .15$ and where about 45% of individuals were included in the first two rows. With $n = 2$ (three-step mediation), $\rho = .3$ yields 28.1%, and $\rho = .5$ yields 40% of the population satisfying ✓✓✓✓ or ⊗ ⊗ ⊗⊗. These results are similar to Table 8, which is based on correlations ranging from .2 to .37, and where about 28% of individuals are tallied in the first two rows.

Table 9 Proportion of the population satisfying ✓✓ . . . ✓ or ⊗⊗ · · · ⊗ in an $(n + 1)$-step mediation

	General case w. constant ρ $(\forall i, \forall j < k, \forall \ell)$ $\rho_{XM_i} = \rho, \rho_{M_jM_k} = \rho, \rho_{M_\ell Y} = \rho$			
n	$\rho = .3$	$\rho = .5$	$\rho = .7$	$\rho = .9$
1	.396	.500	.620	.785
2	.281	.400	.541	.739
3	.209	.333	.486	.706
4	.162	.286	.445	.680
5	.128	.250	.413	.659
6	.105	.222	.387	.642
7	.086	.200	.365	.627
8	.073	.182	.347	.615
9	.062	.167	.331	.604
10	.053	.154	.317	.594
15	.029	.111	.268	.556
20	.018	.087	.236	.531
30	.009	.061	.196	.496

Note. The symbol ✓ denotes a value above the median. All correlations are set equal to a constant ρ

Moving beyond three-step mediations, the table shows a notable decline in the number of people who match the aggregate mediation pattern. When $n > 2$ and $0 < \rho \leq .7$, fewer than half of all individuals align with that pattern. When $n > 3$ and $\rho \leq .5$, fewer than a third of all individuals satisfy ✓✓ . . . ✓ or ⊗ ⊗ · · · ⊗. The table suggests that, in mediations with many variables, the aggregate patterns have very limited scope. Furthermore, in the online supplement, a similar table shows that, when looking at cases of full mediation, the proportions decrease faster as we increase the number of mediating variables.

These insights, even if they are only rough approximations, are of great importance to the design of policies and interventions: It can be of paramount importance to assess beforehand whether there are only very few people who match the profile of the intended target group (e.g., people with low values across all variables, when high values are the desirable outcomes). In the latter case, irrespective of questions of causality associated with the mediation model, an intervention may only apply to few people and/or it may lead to unintended negative consequences for many others (Fairchild & MacKinnon, 2014; MacKinnon, 2011; Pillow et al., 1991).

From medians to other percentiles

We now explore how we can move beyond splitting variables at their median. Rather than defining ✓ and ⊗ as opposite sides of the median, instead, we let ✓ denote either the top or bottom k^{th} percentile, and ⊗ the 'opposite' bottom or top k^{th} percentile. For instance, for the 25^{th} percentile, and a

negative correlation between two variables, we can use ✓✓ to denote the event that the two variables take values in opposite (top and bottom, or bottom and top) quartiles. Regardless of the percentiles of interest, the Appendix provides the mathematics in the section titled "Comparing extreme groups" and our shinyapp allows users to calculate results. Like before, for X, we set ✓ to denote a high value.

Returning, briefly, to the correlation between anxiety and academic performance of our first example, suppose that we are interested in the most dire cases: people in the top-10% on the anxiety scale (who we will call *very anxious*) and in the bottom-10% in academic performance (who we will call *very low performing*). With $\rho = -1$ between anxiety and academic performance, very anxious very low performing individuals would constitute 10% of the population. However, with $\rho = -.15$, individuals who are both very anxious and very low performing fortunately make up a much smaller 1.52% of the population. Taking Owens et al.'s parameters at face value a very similar 1.50% of individuals are very anxious, display below average working memory, and are very low performing. Hence, the conditional probability that a very anxious very low performing person also has below average working memory, is $\frac{150}{152} = .99$. Looking at extremes on all three constructs, 1.05% are very anxious, very low performing and have bottom-10% working memory (Table 10). In other words, the conditional probability that a very anxious, very low performing person also has very low working memory, is $\frac{105}{152} = .69$. These types of insights are important for the design of interventions aimed at improving academic performance: Virtually everybody who is both very anxious and very low performing, also has below average memory. Among people who are both very anxious and very low performing, as many as about two out of three suffer from bottom-10% memory. Looking from a different perspective, however, one should also notice a silver lining, in that, among very anxious individuals with very low working memory (2.66% of the population), only about 39% are also very low performing (1.05% of the population).

Table 10 Proportion of the population satisfying or mismatching the mediation pattern $X \uparrow M \downarrow Y \downarrow$ of Owens et al. (2012) when treating their sample correlations as accurate population correlations

	Mediation Path Links Violated	$X \uparrow$ Anxiety	$M \downarrow$ Working Memory	$Y \downarrow$ Academic Performance	Proportion of the Population
A1	No violation	very high	very low	very low	.0105 .021
A2		very low	very high	very high	.0105
B1	XM	very high	very high	very high	.0009 .0018
B2		very low	very low	very low	.0009
C1	MY	very high	very low	very high	.0000 .0000
C2		very low	very high	very low	.0000
D1	XM; MY	very low	very low	very high	.0000 .0000
D2		very high	very high	very low	.0000
Other					.9772 .9772

Note. Top 10% is labeled as "very high," bottom 10% is labeled as "very low."

The usual view of mediation as a model of population-level relationships leaves out details that are readily available from the joint distribution of the variables of interest. In particular, through the lens of a mediation model, one can cast and evaluate highly detailed and nuanced theories about individual behavior and quantify how many individuals satisfy or violate certain hypotheses. This is possible without having to introduce within-subject mediations and, in particular, without a need for within-subject correlations. Furthermore, as we have already alluded to elsewhere, calculating how many people have a certain combination of features does not require any assumptions about causality. However, in some cases, finding that large numbers of people mismatched a stylized theory, could challenge causal theories if they disallowed such exceptions. In the online supplement, we provide two in-depth illustrations for medians and quartiles on two research paradigms from cognitive neuroscience and applied psychology.

Beyond normal distributions

The use of normal distributions enabled us to derive P_{XM}, P_{MY}, P_{XY} from correlation coefficients via Eq. 1. However, researchers can circumnavigate normal distributions and correlations whenever they have access to empirical data to measure P_{XM}, P_{MY}, P_{XY} directly.[11] Otherwise, we can group mirror-image cases and, regardless of the joint distribution of (X, M, Y), obtain the general result of Table 11. These expressions hold for any distribution when dichotomizing at the median.

Scholars who derive the marginal probabilities P_{XM}, P_{MY}, P_{XY} directly from data may sometimes only be able to obtain these from separate samples of people. In such a case, in addition to ensuring that the samples properly reflect their population of interest (similar medians) and the measures show similar properties in each sample (similar reliabilities), scholars need to check whether there even exists[12] a joint population distribution (joint normal or other) of the variables of interest. As we review in the Appendix, for two-step mediation, with P_{XM}, P_{MY}, P_{XY} inferred from sep-

[11] Recall that P_{XM} is the probability that X and M jointly satisfy ✓✓, whereas P_{MY} is the probability of ✓✓ for M and Y jointly, and P_{XY} denotes the probability of ✓✓ for X and Y jointly.

[12] Conceptually, it should exist, but mathematically, certain probabilities, P_{XM}, P_{MY}, P_{XY}, are incompatible with the existence of a joint distribution.

Table 11 Probability that a randomly drawn individual matches or mismatches a two-step mediation pattern

Mediation Path Links Violated	X	M	Y		X	M	Y	Probability
A No violation	✓	✓	✓	or	⊗	⊗	⊗	$P_{XM} + P_{MY} + P_{XY} - 1/2$
B XM	✓	⊗	⊗	or	⊗	✓	✓	$-P_{XM} + P_{MY} - P_{XY} + 1/2$
C MY	✓	✓	⊗	or	⊗	⊗	✓	$P_{XM} - P_{MY} - P_{XY} + 1/2$
D XM; MY	⊗	✓	⊗	or	✓	⊗	✓	$-P_{XM} - P_{MY} + P_{XY} + 1/2$

Note. These results apply regardless of whether the joint distribution of (X, M, Y) is trivariate normal or not. The symbol ⊗ denotes the opposite of ✓, which, in turn, can denote either above or below median for a given variable. See text for more explanations

arate studies, such a joint distribution exists if and only if the quantities in the right-most column of Table 11 are non-negative. These are the prerequisites, absent a multivariate normal assumption, for running analyses like the ones we have reviewed in our examples when the marginal probabilities have been derived separately and, hence, there is no automatic guarantee that a joint distribution even exists. For three-step mediation, where we consider six marginal probabilities instead of three, the Appendix provides a longer list of properties that need to be satisfied, in order for a joint distribution to exist. It also provides upper and lower bounds on the proportion of people who fully match the mediation pattern without violation. The shinyapp checks these requirements after the user enters correlations of their choice. Note, however, that the shinyapp employs multivariate normal distributions for its calculations.

When we step beyond multivariate normal distributions, we can also consider situations in which the value of one variable moderates the connection between others (see Chmura Kraemer et al., 2008; Edwards & Lambert, 2007; Hayes, 2015; Preacher et al., 2007, for examples). In a two-step mediation, P_{XY} could be a conditional probability that depends on the value of M, or, similarly, P_{XM} could depend on the value of Y. Alternatively, P_{MY} could depend on the value of X. Such interactions among the variables would completely change the calculus behind our tables and we leave this to others to explore. We also note that moderated mediation is subject to some controversy regarding the potential for circular scientific reasoning, such as time or causality loops (see e.g., Chmura Kraemer et al.,2008; Edwards & Lambert, 2007; Hayes, 2017; James & Brett, 1984; Karazsia & Berlin, 2018; MacKinnon, 2011; Muller et al., 2005; Preacher et al., 2007, for related discussions).

Conclusion and discussion

In this paper, we have discussed the model-theory gap in mediation analysis. We established the importance of considering behavior at the individual level across many areas of psychology by providing four in-depth illustrative examples spanning developmental, educational, organizational psychology, and neuroscience (two of them in the online supplemental materials). We have taken the methods, correlation coefficients, measurements, study designs, sample sizes, etc., from these published studies at face value. Every correlation is a population correlation that we assume to accurately capture the covariation of the pertinent variables across the entire population of interest. We consider individuals as realization of the joint random variables that make up the mediation model. An "exception" to a theory is a joint set of values of X, M, Y that mismatches the theory's stylized aggregate pattern.

In Figs. 1 and 3 (and B1 and B2 in the supplemental materials), we reported the percentages of individuals who jointly satisfy each effect of the indirect pathway. In Tables 1-5, 7, 8, and 10 (and B1, B4, B8 and B9 in the supplemental materials), we have listed various ways to match or mismatch the patterns of relationships in mediations and we have provided the proportions of the population who display each match or mismatch. Our illustrative analyses show that mediation results may not fully support the authors' interpretations and can seriously contradict a cursive reader's intuitions. Quite strikingly, we repeatedly find that even with reasonably high correlations, not even half of a population tends to align with the (median-based) pattern of even a simple two-step mediation. That subpopulation is smaller yet for three-step mediations, and it appears to rapidly shrink further with more mediators. However, we do not mean to imply that there is a threshold for the proportion of matches (e.g., half) that would give a stylized theory sufficient scope. Rather, we want to communicate that there is a great disconnect between the stylized aggregate pattern of most mediation models and what actually arises at the individual level (Allport, 1937; Bem & Allen, 1974; Carroll, 2021; Howe et al., 2016; Krull & MacKinnon, 1999; Reise & Widaman, 1999; Witkiewitz et al., 2007).

That disconnect between population-level models and individual-level theory is not merely an esoteric curiosity (Magnusson & Bergman, 1988; Sterba & Bauer, 2010; Carroll, 2021; Collins et al., 1998; Bogat et al., 2016). Users of mediation analysis may naturally wonder how prevalent and how pronounced this disconnect may be in their own research paradigm and in their own data. Our analyses show that, short of near-'perfect' correlations among pairs of variables, rather substantial numbers of individuals mismatch the stylized pattern(s) brought forth by just about any mediation analysis. Although others have considered the differences between aggregate and individual patterns (e.g., Magnusson & Bergman, 1988; von Eye & Bergman, 2003) our results provide a striking demonstration of how this disconnect applies to commonplace analytic paradigms and how the gap between aggregate and individual behavior may be substantially bigger than one might intuitively expect.

At the core, our tabulations provide a concrete picture of the heterogeneity across individuals that a given mediation model assumes and/or implies. One can precisely quantify how few individuals match the stylized aggregate pattern. Just as importantly, one can see, for every type of mismatched pattern, how many people satisfy that pattern, assuming that the mediation model and its correlation structure are perfectly accurate depictions of the entire population. The variable-based approach, in which one examines the strength of a mediated pathway by considering regression coefficients and the products of regression coefficients, provides a far more abstract view of the heterogeneity across individual people[13] (Konstantopoulos et al., 2019; Wiedermann et al., 2022). The person-oriented approach, in turn, while it also uses tabulation of individuals according to shared patterns, is again different from our approach: Its prominent tool, configural frequency analysis (von Eye & Bergman, 2003; von Eye et al., 2009; von Eye & Wiedermann, 2022; Wiedermann & von Eye, 2021; Smyth & MacKinnon, 2021) is primarily aimed at inference, by assessing whether tabulated empirical frequencies significantly differ from what would be expected by chance, or under independence, or under some other null hypothesis.

Unlike our approach, which leverages multivariate normal distributions taken to describe a population, CFA is a data-driven method designed to evaluate whether a group of individuals follows a model. CFA also leads to different interpretations of patterns. Our approach seeks to describe what a given mediation model says about individuals, treating every individual as a sample from said model that could mismatch the stylized theory. On the other hand, CFA focuses on mismatches between observed data and an expected pattern to evaluate a model.

CFA is also useful for comparing different models, for a given sample of data. In contrast, when assuming population correlations, our tabulation does not serve statistical inference. Instead, it helps scholars avoid conjunction fallacies that a variable-level view may trigger, if over-interpreted. For instance, it is worth understanding how many of the individuals who align with the posited XM relationship may misalign with the MY relationship and vice versa. Focusing on just the strength of coefficient products may lead a mediation user to miss these useful details.

By quantifying how many people match the aggregate pattern, our tabulation also helps the scholar evaluate the scope of the stylized theory embodied in that aggregate pattern, and zoom in from the aggregate pattern to the distribution of individual patterns. This provides a platform to dissect the scope and possible unintended consequences of interventions or policy decisions. Again, correlations and regression coefficients do not provide that type and level of resolution to individual behavior. The shinyapp also permits comparisons between two-step and three-step mediations. The existing standard strategy for comparing the strength of two-step vs. three-step mediations - inspecting the drop in the product of regression weights - provides valuable information but leaves much ambiguity. Comparing the tabulation for the two- and three-step mediations provides the scholar with a very precise and concrete account of how the aggregate pattern, as well as all other patterns are distributed in either case, thereby allowing them to better trade off between adding an extra explanatory variable and having fewer individuals aligned with the aggregate pattern.

Applied scientists should inspect the rates of matching and mismatching patterns before offering intervention recommendations. Anyone offering policy advice should evaluate the possible unintended consequences of interventions that might focus on a small minority of individuals who match the pattern, but that could hurt many others. For example, suppose a researcher found that attending math classes worsens anxiety among students and that this decreases academic performance. Hence, the researcher developed a new mathematics course that aims to elicit less anxiety, with the goal of improving performance. However, it may not be effective to simply enroll all students in this course who are suffering academically. The course is only designed to improve performance in students who both 1) show math-related anxiety, and 2) suffer academically due to anxiety. The proportion of individuals satisfying both criteria may be limited.

The ideas here apply regardless of whether we aim to help anxious people perform better, consider interventions to support psychological growth over the life span, aim to enhance people's health, or boost leadership effectiveness. For each of these cases, tabulating matches and mismatches of stylized verbal descriptions of the mediation pathways is key to weighing and balancing the positive and negative impacts of policies. We only peripherally touched on causality issues: Mapping out how many people match which profile does not depend on causal interpretations, but finding that large numbers of individuals are an exception to a stylized theory can call causal interpretations of that theory into question.

How do the insights in this paper connect with other major efforts, under way across Psychology, to enhance the scientific value of our discipline? In contemporary psychological science, much of the discussion about trustworthiness and quality of psychological research is dominated by issues of replicability. Our approach suggests that scholars who run a replication study of a mediation model may want to compare the tables of matching and mismatching pattern probabilities between the two studies beyond just comparing the qualitative pattern of correlations. When replicating old studies that only reported correlations, new studies can go further

[13] Note that some emerging variable-based methods, which use regressions but model distributions beyond just the average, also track some individual-level phenomena

and directly measure the numbers of people who match or mismatch the stylized aggregate mediation pattern. Whether one considers a replication to be successful, all the way to the level of distributions over individual people, can be assessed with both person-oriented methods and traditional aggregate-level techniques. In addition, we would warn that, just like logically incorrect causality interpretations cannot be detected through replication, so is replication blind to other reasoning errors: Mediation models that solely generate population-level results cannot adequately address a question about individual behavior, regardless of whether results replicate.

In this paper, we warn of a specific set of scientific reasoning fallacies that can occur when working with mediations. We have discussed a potential logical fallacy of sweeping generalization from mediation analyses of data pooled across individuals. Scholars should take great care not to over-interpret stylized verbal descriptions of mediations as if they applied to all individuals. Many paradigms are plagued with a disconnect between aggregate and individual behavior. Fortunately, mediation models describe the covariation of people in a population of interest and provide readily available information about how many people match or mismatch a particular pattern. We have provided examples and some theoretical results here. The Appendix and supplemental materials provide numerous technical results. An online tool at https://herulor.shinyapps.io/MediationApp/ makes it easy for scholars to carry out similar analyses for two-step and three-step mediations of their choice. The app allows the user to enter correlations and percentiles of interest, and, given the correlations are mutually compatible, it will generate a table similar to the tables in this paper. In all, mediation is a paradigm in which we can readily overcome and avoid certain fallacies of sweeping generalization.

Because we are interested in the theoretical scope of well-established results, we have taken the study correlations at face value, as though they represented fully accurate population correlations. To the extent that the sample correlations in the illustrative studies were reasonable point estimates of the corresponding population correlations, so are our percentages, proportions, or probabilities in the figures and tables reasonable point estimates. In an effort to move from population distributions to empirical data, the shinyapp also allows users to enter sample correlations instead of population correlations. In that case, the shinyapp adds some basic statistical inference, in the form of marginal confidence intervals, to the tabulation. Table B1 in the online supplement provides an example.

Future work could develop additional tools to better accommodate the uncertainty that is inherent in inferring correlations from finite samples of data. Such tools may thereby quantify uncertainty about the inferred pattern probabilities in our figures and tables. Relatedly, while our dichotomization-based approach (splitting participants at the median) permitted clear interpretation and intuition, examining these notions without dichotomization may be also valuable. For instance, considering extreme cases implicitly introduced three categories of participants. One could naturally extend to more categories, especially when assuming multivariate normality. Alternatively, when using dichotomization, one could split the variables at a percentile other than at the median. These dichotomizations would lead to expressions that, while different from the ones we provided here, would be of a similar mathematical form, as long as each variable is dichotomized at the same percentile. Future work could also consider various generalizations and enhancements, such as deriving more tools for understanding mediations with more than three steps and exploring how measured variables can fall short in tracking the constructs of interest. Likewise, future work could expand this approach to related models such as structural equation models.

Because this is theoretical work, we have idealized each individual as having a single value for each variable. In actual data, we face variability, not just between individuals, but also within individuals. Variation in repeated observations within a single person can be due to a combination of varying hypothetical constructs, as well as probabilistic measurement errors. From a theoretical viewpoint, if a mediation model describes the covariation of variables within a single individual, then the type of analyses we have discussed captures, not how many people match or mismatch the pattern revealed by that model, but rather the proportion of time that this given individual matches or mismatches his or her aggregate mediation pattern. Translating the lessons we learned from our analysis above to the variation within an individual, this also means that, as the number of variables in a within-person mediation model goes up, the individual satisfies the pattern of that mediation more and more rarely. We expect these combinations of insights to also apply in multi-level mediation (Bauer et al., 2006; Preacher & Selig, 2012), where the modeling directly addresses ergodicity (or lack thereof).

Moving from a purely theoretical perspective to inference from data, in a fashion that combines these sources of variability with measurement error, furthermore raises many interesting methodological questions. We leave these for future work.

Last, but not least, we need to raise concerns for studies that use a very large number of mediators. Interpreting very large and complicated mediation models at the aggregate level is almost certainly disconnected from the individual behavior of almost every person. On the surface, large mediation models would seem like particularly refined, detailed, and nuanced theories. However, Saylors & Trafimow (2020) warned that it may be nearly impossible to infer the correct mediation structure from data (see also Trafimow, 2017). Suppose that we disregard their warnings and we assume, as

we have throughout this paper, that we have inferred a correct mediation and fully valid population correlations. Unless all population correlations are extremely high, only few individual people will actually behave in a way that aligns with that stylized aggregate pattern. For large numbers of mediator variables, most people, if not almost everybody, will be an exception to the verbal story attached to that mediation. In these cases, mediation may be a good theory of aggregate patterns of behavior and therefore of value to sociologists or political scientists. However, policy makers will still need to be extremely careful in how they design interventions to ensure that they are applicable to more than a few people. In all, for a very large mediation model, when almost everybody is an exception to its aggregate pattern of relationships among variables, that stylized mediation pattern offers questionable value as a theory of individual people.

Open Practices Statement We did not collect any data from participants. Instead, we re-analyzed findings from published manuscripts. Citations for those manuscripts can be found nearby their corresponding re-analysis. Our results can be reproduced using the released open-access shinyapp https://herulor.shinyapps.io/MediationApp/. The R code for the shinyapp can be found at https://github.com/herulor/MediationApp.

Appendix A: Technical results

A.1 Two-step mediation

The main text presents most technical results for two-step mediation. Below, we present the details regarding the existence of a joint probability distribution.

A.1.1 On the existence of a joint probability distribution

Note that any values P_{XM}, P_{MY}, and P_{XY}, that are consistent with jointly distributed $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$, will satisfy the four inequalities

$$P_{XM} + P_{MY} + P_{XY} - 1/2 \geq 0, \tag{10}$$
$$-P_{XM} + P_{MY} - P_{XY} + 1/2 \geq 0, \tag{11}$$
$$P_{XM} - P_{MY} - P_{XY} + 1/2 \geq 0, \tag{12}$$
$$-P_{XM} - P_{MY} + P_{XY} + 1/2 \geq 0, \tag{13}$$

as long as the initial distribution of the variables X, M, and Y can be dichotomized uniformly at the median.

Moreover, if the probabilities P_{XM}, P_{MY}, and P_{XY} are directly computed from data that have been obtained from separate samples of people, then inequalities 10-13 are also sufficient for the existence of a joint distribution of $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ (see, e.g., Suppes & Zanotti, 1981, p.198). Note that some combinations of the probabilities P_{XM}, P_{MY}, and P_{XY} obtained from separate samples may be compatible with each other as described by these inequalities, but still incompatible with a trivariate normal distribution. In situations when the scholar assumes trivariate normality, it may be simpler to verify, instead, that the correlation matrix created from the separately estimated ρ_{XM}, ρ_{MY}, and ρ_{XY} is positive (semi)definite. In the shinyapp, the probabilities P_{XM}, P_{MY}, and P_{XY} are computed assuming normal distributions. Thus, the app verifies that the input correlations produce a positive (semi)definite matrix.

A.2 Three-step mediation

We shall assume that the variables X, M_1, M_2, and Y are jointly distributed continuous random variables such that they can be uniformly dichotomized at the median. Let ρ_{AB}, with $A, B \in \{X, M_1, M_2, Y\}$, denote the correlation coefficient for each pair of variables. Analogously to the previous section, define $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ as the corresponding dichotomized variables taking the values ✓ and ⊗ if the original variable took a value above or below the median as explained below. Writing $x_{50}, m_{1,50}, m_{2,50}, y_{50}$ to denote the medians of X, M_1, M_2, and Y, respectively, let

$$\begin{aligned}
\widetilde{X} = ✓ &\iff X \geq x_{50},\\
\widetilde{X} = ⊗ &\iff X < x_{50},\\
\widetilde{M}_1 = ✓ &\iff \operatorname{sign}(\rho_{XM_1}) M_1 \geq m_{1,50},\\
\widetilde{M}_1 = ⊗ &\iff \operatorname{sign}(\rho_{XM_1}) M_1 < m_{1,50},\\
\widetilde{M}_2 = ✓ &\iff \operatorname{sign}(\rho_{XM_1}\rho_{M_1M_2}) M_2 \geq m_{2,50},\\
\widetilde{M}_2 = ⊗ &\iff \operatorname{sign}(\rho_{XM_1}\rho_{M_1M_2}) M_2 < m_{2,50},\\
\widetilde{Y} = ✓ &\iff \operatorname{sign}(\rho_{XM}\rho_{M_1M_2}\rho_{M_2Y}) Y \geq y_{50},\\
\widetilde{Y} = ⊗ &\iff \operatorname{sign}(\rho_{XM}\rho_{M_1M_2}\rho_{M_2Y}) Y < y_{50}.
\end{aligned}$$

We describe the pairwise joint distributions of the dichotomizations in a manner similar to the bivariate distributions for the two-step mediation. The pairwise joint distributions of the dichotomizations are given in the contingency tables in Fig. 4. If we assume that X, M_1, M_2, and Y have pairwise bivariate normal distributions, we can compute the probabilities P_{XM_1}, P_{M_1Y}, $P_{M_1M_2}$, P_{XM_2}, P_{M_2Y}, and P_{XY} as

$$P_{XM_1} = \frac{1}{4} + \frac{\arcsin\left(\left|\rho_{XM_1}\right|\right)}{2\pi},$$
$$P_{M_1M_2} = \frac{1}{4} + \frac{\arcsin\left(\left|\rho_{M_1M_2}\right|\right)}{2\pi},$$
$$P_{M_2Y} = \frac{1}{4} + \frac{\arcsin\left(\left|\rho_{M_2Y}\right|\right)}{2\pi},$$

	$\widetilde{M}_1 = \checkmark$	$\widetilde{M}_1 = \otimes$	
$\widetilde{X} = \checkmark$	P_{XM_1}	$1/2 - P_{XM_1}$	$1/2$
$\widetilde{X} = \otimes$	$1/2 - P_{XM_1}$	P_{XM_1}	$1/2$
	$1/2$	$1/2$	

	$\widetilde{M}_2 = \checkmark$	$\widetilde{M}_2 = \otimes$	
$\widetilde{X} = \checkmark$	P_{XM_2}	$1/2 - P_{XM_2}$	$1/2$
$\widetilde{X} = \otimes$	$1/2 - P_{XM_2}$	P_{XM_2}	$1/2$
	$1/2$	$1/2$	

	$\widetilde{Y} = \checkmark$	$\widetilde{Y} = \otimes$	
$\widetilde{M}_1 = \checkmark$	P_{M_1Y}	$1/2 - P_{M_1Y}$	$1/2$
$\widetilde{M}_1 = \otimes$	$1/2 - P_{M_1Y}$	P_{M_1Y}	$1/2$
	$1/2$	$1/2$	

	$\widetilde{Y} = \checkmark$	$\widetilde{Y} = \otimes$	
$\widetilde{M}_2 = \checkmark$	P_{M_2Y}	$1/2 - P_{M_2Y}$	$1/2$
$\widetilde{M}_2 = \otimes$	$1/2 - P_{M_2Y}$	P_{M_2Y}	$1/2$
	$1/2$	$1/2$	

	$\widetilde{M}_2 = \checkmark$	$\widetilde{M}_2 = \otimes$	
$\widetilde{M}_1 = \checkmark$	$P_{M_1M_2}$	$1/2 - P_{M_1M_2}$	$1/2$
$\widetilde{M}_1 = \otimes$	$1/2 - P_{M_1M_2}$	$P_{M_1M_2}$	$1/2$
	$1/2$	$1/2$	

	$\widetilde{Y} = \checkmark$	$\widetilde{Y} = \otimes$	
$\widetilde{X} = \checkmark$	P_{XY}	$1/2 - P_{XY}$	$1/2$
$\widetilde{X} = \otimes$	$1/2 - P_{XY}$	P_{XY}	$1/2$
	$1/2$	$1/2$	

Fig. 4 Pairwise joint distributions of variables $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$

$$P_{XM_2} = \frac{1}{4} + \frac{\arcsin\left(\text{sign}(\rho_{XM_1}\rho_{M_1M_2})\rho_{XM_2}\right)}{2\pi},$$
$$P_{M_1Y} = \frac{1}{4} + \frac{\arcsin\left(\text{sign}(\rho_{M_1M_2}\rho_{M_2Y})\rho_{M_1Y}\right)}{2\pi},$$
$$P_{XY} = \frac{1}{4} + \frac{\arcsin\left(\text{sign}(\rho_{XM_1}\rho_{M_1M_2}\rho_{M_2Y})\rho_{XY}\right)}{2\pi}.$$

The joint distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ is characterized by the probabilities of the 16 possible combinations of values that $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ can take, as shown in Table 12. This distribution is not entirely determined by the pairwise probabilities identified above. Now, we will derive the properties of the joint distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ that can be deduced only from the pairwise probabilities. First, note that each subset of three variables of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ can be analyzed as a two-step mediation. From the analysis in the previous section, we know that the pairwise probabilities determine the probabilities of some events of each of these triples of random variables. In particular, the probability with which all three variables in any of these triples are equal to each other is determined. There are four of these triples, hence there are four such probabilities.

We define, for instance, the following probability.

$$P_{XM_1Y} = \Pr(\widetilde{X}=\checkmark, \widetilde{M}_1=\checkmark, \widetilde{Y}=\checkmark) + \Pr(\widetilde{X}=\otimes, \widetilde{M}_1=\otimes, \widetilde{Y}=\otimes).$$

The other three probabilities are labeled analogously. The four probabilities of equality of triples of random variables are given by the following expressions.

$$P_{XM_1M_2} = P_{XM_1} + P_{M_1M_2} + P_{XM_2} - 1/2, \quad (14)$$
$$P_{XM_1Y} = P_{XM_1} + P_{M_1Y} + P_{XY} - 1/2, \quad (15)$$
$$P_{XM_2Y} = P_{XM_2} + P_{M_2Y} + P_{XY} - 1/2, \quad (16)$$
$$P_{M_1M_2Y} = P_{M_1M_2} + P_{M_2Y} + P_{M_1Y} - 1/2. \quad (17)$$

The 16 probabilities $\Pr_{A1}, \ldots, \Pr_{H2}$ are real numbers between 0 and 1. They sum to 1, and satisfy the following sets of linear constraints:

a) The margins for each variable are equal to 1/2.

$$\Pr_{A1} + \Pr_{B1} + \Pr_{C2} + \Pr_{D2} + \Pr_{E2} + \Pr_{F1} + \Pr_{G2} + \Pr_{H1} = \Pr(\widetilde{X} = \checkmark) = 1/2, \quad (18)$$

$$\Pr_{A1} + \Pr_{B2} + \Pr_{C1} + \Pr_{D2} + \Pr_{E2} + \Pr_{F1} + \Pr_{G1} + \Pr_{H2} = \Pr(\widetilde{M}_1 = \checkmark) = 1/2, \quad (19)$$

$$\Pr_{A1} + \Pr_{B2} + \Pr_{C2} + \Pr_{D1} + \Pr_{E2} + \Pr_{F2} + \Pr_{G1} + \Pr_{H1} = \Pr(\widetilde{M}_2 = \checkmark) = 1/2, \quad (20)$$

$$\Pr_{A1} + \Pr_{B2} + \Pr_{C2} + \Pr_{D2} + \Pr_{E1} + \Pr_{F2} + \Pr_{G2} + \Pr_{H2} = \Pr(\widetilde{Y} = \checkmark) = 1/2. \quad (21)$$

b) The probabilities associated with the pairwise correlations are known.

$$\Pr_{A1} + \Pr_{D2} + \Pr_{E2} + \Pr_{F1} = P_{XM_1}, \quad (22)$$
$$\Pr_{A1} + \Pr_{B2} + \Pr_{D2} + \Pr_{H2} = P_{M_1Y}, \quad (23)$$
$$\Pr_{A1} + \Pr_{C2} + \Pr_{D2} + \Pr_{G2} = P_{XY}, \quad (24)$$

Table 12 Distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$

	$\widetilde{X}$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{Y}$	Joint Outcome Probability
$A1$	✓	✓	✓	✓	$\Pr_{A1}$
$A2$	⊗	⊗	⊗	⊗	$\Pr_{A2}$
$B1$	✓	⊗	⊗	⊗	$\Pr_{B1}$
$B2$	⊗	✓	✓	✓	$\Pr_{B2}$
$C1$	⊗	✓	⊗	⊗	$\Pr_{C1}$
$C2$	✓	⊗	✓	✓	$\Pr_{C2}$
$D1$	⊗	⊗	✓	⊗	$\Pr_{D1}$
$D2$	✓	✓	⊗	✓	$\Pr_{D2}$
$E1$	⊗	⊗	⊗	✓	$\Pr_{E1}$
$E2$	✓	✓	✓	⊗	$\Pr_{E2}$
$F1$	✓	✓	⊗	⊗	$\Pr_{F1}$
$F2$	⊗	⊗	✓	✓	$\Pr_{F2}$
$G1$	⊗	✓	✓	⊗	$\Pr_{G1}$
$G2$	✓	⊗	⊗	✓	$\Pr_{G2}$
$H1$	✓	⊗	✓	⊗	$\Pr_{H1}$
$H2$	⊗	✓	⊗	✓	$\Pr_{H2}$

$$\Pr_{A1} + \Pr_{B2} + \Pr_{E2} + \Pr_{G1} = P_{M_1M_2}, \tag{25}$$
$$\Pr_{A1} + \Pr_{C2} + \Pr_{E2} + \Pr_{H1} = P_{XM_2}, \tag{26}$$
$$\Pr_{A1} + \Pr_{B2} + \Pr_{C2} + \Pr_{F2} = P_{M_2Y}. \tag{27}$$

If we additionally assume that any two mutually opposite patterns of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ have equal probabilities, then the following four linear constraints must also hold.

$$\Pr_{A1} + \Pr_{E2} = \Pr(\widetilde{X}=✓, \widetilde{M}_1=✓, \widetilde{M}_2=✓) = \frac{1}{2}P_{XM_1M_2}, \tag{28}$$
$$\Pr_{A1} + \Pr_{D2} = \Pr(\widetilde{X}=✓, \widetilde{M}_1=✓, \widetilde{Y}=✓) = \frac{1}{2}P_{XM_1Y}, \tag{29}$$
$$\Pr_{A1} + \Pr_{C2} = \Pr(\widetilde{X}=✓, \widetilde{M}_2=✓, \widetilde{Y}=✓) = \frac{1}{2}P_{XM_2Y}, \tag{30}$$
$$\Pr_{A1} + \Pr_{B2} = \Pr(\widetilde{M}_1=✓, \widetilde{M}_2=✓, \widetilde{Y}=✓) = \frac{1}{2}P_{M_1M_2Y}. \tag{31}$$

Equations 18–31, together with the requirement that the 16 values sum to 1, constitute 15 linearly independent constraints on the probabilities governing the joint distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$ and $\widetilde{Y}$. The constraints leave a single degree of freedom to determine the values of the 16 probabilities. We assign the value a arbitrarily to $\Pr_{A1}$. Then, the remaining probabilities are completely determined by the constraints. For example, $\Pr_{E2} = \frac{1}{2}P_{XM_1M_2} - a$. This yields the distribution in Table 13.

This characterization of the distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$, and $\widetilde{Y}$ implies that for any set of jointly distributed variables X, M_1, M_2, and Y (which can be uniformly dichotomized by their medians and satisfy that opposite patterns of those dichotomizations have equal probabilities), the proportion of the population that matches the full mediation pattern is bounded.

This proportion equals $\Pr_{A1} + \Pr_{A2} = 2a$ and it is bounded from below by

$$\begin{aligned}\max(0,\ & P_{XM_2} + P_{M_1M_2} + P_{M_1Y} + P_{XY} - 1,\\ & P_{XM_1} + P_{M_1Y} + P_{M_2Y} + P_{XM_2} - 1,\\ & P_{XM_1} + P_{M_1M_2} + P_{M_2Y} + P_{XY} - 1).\end{aligned}$$

It is bounded from above by

$$\min(P_{XM_1M_2}, P_{XM_1Y}, P_{XM_2Y}, P_{M_1M_2Y}).$$

If we assume multivariate normality of X, M_1, M_2, and Y, then the probabilities $\Pr_{A1}, \ldots, \Pr_{H2}$ are completely determined. They can be computed using algorithms to find Gaussian multivariate integrals over hyper-rectangular regions, such as those implemented in the R package `mvtnorm` (Genz et al., 2020). The shinyapp uses this package to perform all computations involving probabilities of multivariate normal distributions.

A.2.1 On the existence of a joint distribution

Suppose that we estimate the values P_{XM_1}, P_{M_1Y}, $P_{M_1M_2}$, P_{XM_2}, P_{M_2Y}, and P_{XY} on separate samples. Then the following sets of inequalities (see, Dzhafarov & Kujala, 2016) must hold in order for a joint distribution of the four variables X, M_1, M_2, and Y to exist.

For each triple

$$\begin{aligned}(p_1, p_2, p_3) \in \{&(P_{XM_1}, P_{M_1Y}, P_{XY}),\\ &(P_{XM_1}, P_{M_1M_2}, P_{XM_2}),\\ &(P_{XM_2}, P_{M_2Y}, P_{XY}),\\ &(P_{M_1M_2}, P_{M_2Y}, P_{M_1Y})\},\end{aligned}$$

the following four inequalities need to hold (Suppes & Zanotti, 1981; Araújo et al., 2013).

$$p_1 + p_2 + p_3 - 1/2 \geq 0,$$
$$p_1 - p_2 - p_3 + 1/2 \geq 0,$$

Table 13 Probability values for the distribution of $\widetilde{X}$, $\widetilde{M}_1$, $\widetilde{M}_2$ and $\widetilde{Y}$ with symmetry assumption

	$\widetilde{X}$	$\widetilde{M}_1$	$\widetilde{M}_2$	$\widetilde{Y}$	Joint Outcome Probability
$A1$	✓	✓	✓	✓	$\Pr_{A1} = a$
$A2$	⊗	⊗	⊗	⊗	$\Pr_{A2} = a$
$B1$	✓	⊗	⊗	⊗	$\Pr_{B1} = \frac{1}{2}(P_{M_1M_2} + P_{M_2Y} + P_{M_1Y} - 1/2) - a$
$B2$	⊗	✓	✓	✓	$\Pr_{B2} = \frac{1}{2}(P_{M_1M_2} + P_{M_2Y} + P_{M_1Y} - 1/2) - a$
$C1$	⊗	✓	⊗	⊗	$\Pr_{C1} = \frac{1}{2}(P_{XM_2} + P_{M_2Y} + P_{XY} - 1/2) - a$
$C2$	✓	⊗	✓	✓	$\Pr_{C2} = \frac{1}{2}(P_{XM_2} + P_{M_2Y} + P_{XY} - 1/2) - a$
$D1$	⊗	⊗	✓	⊗	$\Pr_{D1} = \frac{1}{2}(P_{XM_1} + P_{M_1Y} + P_{XY} - 1/2) - a$
$D2$	✓	✓	⊗	✓	$\Pr_{D2} = \frac{1}{2}(P_{XM_1} + P_{M_1Y} + P_{XY} - 1/2) - a$
$E1$	⊗	⊗	⊗	✓	$\Pr_{E1} = \frac{1}{2}(P_{XM_1} + P_{M_1M_2} + P_{XM_2} - 1/2) - a$
$E2$	✓	✓	✓	⊗	$\Pr_{E2} = \frac{1}{2}(P_{XM_1} + P_{M_1M_2} + P_{XM_2} - 1/2) - a$
$F1$	✓	✓	⊗	⊗	$\Pr_{F1} = \frac{1}{2}(1 - P_{XM_2} - P_{M_1M_2} - P_{M_1Y} - P_{XY}) + a$
$F2$	⊗	⊗	✓	✓	$\Pr_{F2} = \frac{1}{2}(1 - P_{XM_2} - P_{M_1M_2} - P_{M_1Y} - P_{XY}) + a$
$G1$	⊗	✓	✓	⊗	$\Pr_{G1} = \frac{1}{2}(1 - P_{XM_1} - P_{M_1Y} - P_{M_2Y} - P_{XM_2}) + a$
$G2$	✓	⊗	⊗	✓	$\Pr_{G2} = \frac{1}{2}(1 - P_{XM_1} - P_{M_1Y} - P_{M_2Y} - P_{XM_2}) + a$
$H1$	✓	⊗	✓	⊗	$\Pr_{H1} = \frac{1}{2}(1 - P_{XM_1} - P_{M_1M_2} - P_{M_2Y} - P_{XY}) + a$
$H2$	⊗	✓	⊗	✓	$\Pr_{H2} = \frac{1}{2}(1 - P_{XM_1} - P_{M_1M_2} - P_{M_2Y} - P_{XY}) + a$

$$-p_1 + p_2 - p_3 + 1/2 \geq 0,$$
$$-p_1 - p_2 + p_3 + 1/2 \geq 0.$$

For each quadruple

$$(p_1, p_2, p_3, p_4) \in \{(P_{XM_1}, P_{M_1M_2}, P_{M_2Y}, P_{XY}), (P_{XM_1}, P_{M_1Y}, P_{M_2Y}, P_{XM_2}), (P_{XM_2}, P_{M_1M_2}, P_{M_1Y}, P_{XY})\},$$

the following eight inequalities need to hold (Clauser et al., 1969; Araújo et al., 2013).

$$\begin{aligned}
p_1 + p_2 + p_3 - p_4 + 1 &\geq 0,\\
-p_1 - p_2 - p_3 + p_4 + 1 &\geq 0,\\
p_1 + p_2 - p_3 + p_4 + 1 &\geq 0,\\
-p_1 - p_2 + p_3 - p_4 + 1 &\geq 0,\\
p_1 - p_2 + p_3 + p_4 + 1 &\geq 0,\\
-p_1 + p_2 - p_3 - p_4 + 1 &\geq 0,\\
-p_1 + p_2 + p_3 + p_4 + 1 &\geq 0,\\
p_1 - p_2 - p_3 - p_4 + 1 &\geq 0.
\end{aligned}$$

Similarly to the two-step case, if the researcher assumes multivariate normal distribution, it may be simpler to just verify that the correlation matrix constructed from the separately estimated pairwise correlations is positive (semi)definite. The shinyapp computes the probabilities P_{XM_1}, P_{M_1Y}, $P_{M_1M_2}$, P_{XM_2}, P_{M_2Y}, and P_{XY} assuming normal distributions. Thus, the app verifies that the corresponding matrix is positive (semi)definite.

A.3 Comparing extreme groups

We now consider extreme groups defined by the top and bottom $100\alpha^{\text{th}}$ percentile. For instance, when $\alpha = .10$, this gives the top 10% together with the bottom 10%. This means that we effectively categorize X, M, and Y into three categories each: the bottom 100α%, non-extreme values, and the top 100α%. We encode that categorization as ✓, −, ⊗. We follow the same procedure as before, when splitting at the median, to determine whether we label a top group or a bottom group as ✓. The pairwise joint distributions of these categorizations are given in the tables in Fig. 5.

The joint distribution of the three categorized variables $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ is determined by the 27 probabilities of the joint outcomes $\tilde{x}\tilde{m}\tilde{y}$, where $\tilde{x}, \tilde{m}, \tilde{y} \in \{✓, -, ⊗\}$. We denote

	$\widetilde{M}=\checkmark$	$\widetilde{M}=-$	$\widetilde{M}=\otimes$	
$\widetilde{X}=\checkmark$	$P_{XM,1}$	$\alpha - P_{XM,1} - P_{XM,2}$	$P_{XM,2}$	α
$\widetilde{X}=-$	$\alpha - P_{XM,1} - P_{XM,3}$	$1-4\alpha+\sum_{i=1}^{4} P_{XM,i}$	$\alpha - P_{XM,2} - P_{XM,4}$	$1-2\alpha$
$\widetilde{X}=\otimes$	$P_{XM,3}$	$\alpha - P_{XM,3} - P_{XM,4}$	$P_{XM,4}$	α
	α	$1-2\alpha$	α	

	$\widetilde{Y}=\checkmark$	$\widetilde{Y}=-$	$\widetilde{Y}=\otimes$	
$\widetilde{M}=\checkmark$	$P_{MY,1}$	$\alpha - P_{MY,1} - P_{MY,2}$	$P_{MY,2}$	α
$\widetilde{M}=-$	$\alpha - P_{MY,1} - P_{MY,3}$	$1-4\alpha+\sum_{i=1}^{4} P_{MY,i}$	$\alpha - P_{MY,2} - P_{MY,4}$	$1-2\alpha$
$\widetilde{M}=\otimes$	$P_{MY,3}$	$\alpha - P_{MY,3} - P_{MY,4}$	$P_{MY,4}$	α
	α	$1-2\alpha$	α	

	$\widetilde{Y}=\checkmark$	$\widetilde{Y}=-$	$\widetilde{Y}=\otimes$	
$\widetilde{X}=\checkmark$	$P_{XY,1}$	$\alpha - P_{XY,1} - P_{XY,2}$	$P_{XY,2}$	α
$\widetilde{X}=-$	$\alpha - P_{XY,1} - P_{XY,3}$	$1-4\alpha+\sum_{i=1}^{4} P_{XY,i}$	$\alpha - P_{XY,2} - P_{XY,4}$	$1-2\alpha$
$\widetilde{X}=\otimes$	$P_{XY,3}$	$\alpha - P_{XY,3} - P_{XY,4}$	$P_{XY,4}$	α
	α	$1-2\alpha$	α	

Fig. 5 Pairwise joint distributions of variables $\widetilde{X}$, $\widetilde{M}$, and $\widetilde{Y}$ in the comparison of extreme groups

the eight extreme group probabilities in the same manner as when we considered median dichotomization.

That is, for instance,

$$\Pr_{A1} = \Pr(\widetilde{X}=\checkmark, \widetilde{M}=\checkmark, \widetilde{Y}=\checkmark),$$
$$\Pr_{B1} = \Pr(\widetilde{X}=\checkmark, \widetilde{M}=\otimes, \widetilde{Y}=\otimes),$$

and similarly for probabilities $\Pr_{A2}$, $\Pr_{B2}$, $\Pr_{C1}$, $\Pr_{C2}$, $\Pr_{D1}$, and $\Pr_{D2}$, as shown in Table 2. We noted that the distribution of the variables obtained by dichotomizing at the medians was completely determined by the bivariate distributions under the symmetry assumption $\Pr_{A1} = \Pr_{A2}$. The joint distribution of the categorization for extreme groups, however, is not completely determined in the same way by the bivariate distributions in Fig. 5, even if we added a symmetry assumption. In particular, neither the probabilities of the extreme groups nor the sum of probabilities of opposite patterns (such as probabilities $\Pr_{A1}$ and $\Pr_{A2}$ for patterns $\checkmark\checkmark\checkmark$ and $\otimes\otimes\otimes$) are completely determined by the bivariate distributions.

Without additional assumptions, inequalities 32–52 characterize the set of values that the vector of probabilities $(\Pr_{A1}, \Pr_{A2}, \Pr_{B1}, \Pr_{B2}, \Pr_{C1}, \Pr_{C2}, \Pr_{D1}, \Pr_{D2})$ can take.

$$\Pr_{A1} \le \min\{P_{XM,1}, P_{MY,1}, P_{XY,1}\}, \tag{32}$$
$$\Pr_{A2} \le \min\{P_{XM,4}, P_{MY,4}, P_{XY,4}\}, \tag{33}$$
$$\Pr_{B1} \le \min\{P_{XM,2}, P_{MY,4}, P_{XY,2}\}, \tag{34}$$
$$\Pr_{B2} \le \min\{P_{XM,3}, P_{MY,1}, P_{XY,3}\}, \tag{35}$$
$$\Pr_{C1} \le \min\{P_{XM,1}, P_{MY,2}, P_{XY,2}\}, \tag{36}$$
$$\Pr_{C2} \le \min\{P_{XM,4}, P_{MY,3}, P_{XY,3}\}, \tag{37}$$
$$\Pr_{D1} \le \min\{P_{XM,2}, P_{MY,3}, P_{XY,1}\}, \tag{38}$$
$$\Pr_{D2} \le \min\{P_{XM,3}, P_{MY,2}, P_{XY,4}\}, \tag{39}$$

$$\Pr_{A1} + \Pr_{C1} \le P_{XM,1}, \tag{40}$$
$$\Pr_{B1} + \Pr_{D1} \le P_{XM,2}, \tag{41}$$
$$\Pr_{B2} + \Pr_{D2} \le P_{XM,3}, \tag{42}$$
$$\Pr_{A2} + \Pr_{C2} \le P_{XM,4}, \tag{43}$$

$$\Pr_{A1} + \Pr_{B2} \le P_{MY,1}, \tag{44}$$
$$\Pr_{C1} + \Pr_{D2} \le P_{MY,2}, \tag{45}$$
$$\Pr_{C2} + \Pr_{D2} \le P_{MY,3}, \tag{46}$$
$$\Pr_{A2} + \Pr_{B1} \le P_{MY,4}, \tag{47}$$

$$\Pr_{A1} + \Pr_{D1} \leq P_{XY,1}, \quad (48)$$
$$\Pr_{B1} + \Pr_{C1} \leq P_{XY,2}, \quad (49)$$
$$\Pr_{B2} + \Pr_{C2} \leq P_{XY,3}, \quad (50)$$
$$\Pr_{A2} + \Pr_{D2} \leq P_{XY,4}, \quad (51)$$

$$\begin{aligned}\Pr_{A1}+\Pr_{A2}+\Pr_{B1}+\Pr_{B2}+\Pr_{C1}+\Pr_{C2}+\Pr_{D1}+\Pr_{D2} \leq \\ \min\{P_{XM,1}+P_{XM,2}+P_{XM,3}+P_{XM,4}, \\ P_{MY,1}+P_{MY,2}+P_{MY,3}+P_{MY,4}, \\ P_{XY,1}+P_{XY,2}+P_{XY,3}+P_{XY,4}\} \leq 2\alpha. \quad (52)\end{aligned}$$

The inequalities 32–52 hold for any set of random variables X, M, and Y, categorized into the bottom $100\alpha\%$, non-extreme values, and the top $100\alpha\%$. We need more assumptions to uniquely identify values of the probabilities $\Pr_{A1}$ $\Pr_{A2}$, $\Pr_{B1}$, $\Pr_{B2}$, $\Pr_{C1}$, $\Pr_{C2}$, $\Pr_{D1}$, and $\Pr_{D2}$.

For instance, if we assume trivariate normality of X, M, and$Y$, we can compute these probabilities using algorithms for Gaussian multivariate integrals over hyper-rectangular regions, such as those implemented in the R package `mvtnorm` (Genz et al., 2020). For computations involving extreme groups, the shinyapp uses this package.

Moreover, under trivariate normality, we can simplify the tables in Fig. 5 and compute the probabilities defined in them based on the pairwise correlations ρ_{XM}, ρ_{MY}, and ρ_{XY}. First, the symmetry of the pairwise probabilities under bivariate normal distributions, implies that $P_{XM,1} = P_{XM,4}$, $P_{XM,2} = P_{XM,3}$, and similarly for the corresponding pairs of $P_{MY,i}$ and $P_{XY,i}$. Thus, we can simplify the description of the pairwise distributions letting $P_{XM} = P_{XM,1}$, $P'_{XM} = P_{XM,2}$, and, analogously, define P_{MY}, P'_{MY}, P_{XY}, and P'_{XY}. The proportions P_{XM}, P'_{XM}, P_{MY}, P'_{MY}, P_{XY}, and P'_{XY} can then be computed (Owen, 1980) using

$$\alpha - 2\phi\left(\Phi^{-1}(\alpha)\right)\int_0^{h(\rho)} \frac{\phi\left(\Phi^{-1}(\alpha)t\right)}{1+t^2}\,dt,$$

where ϕ, and Φ denote the density and the distribution function of the standard normal, respectively, and where

$$h(\rho) = \sqrt{\frac{1-|\rho_{XM}|}{1+|\rho_{XM}|}} \text{ for } P_{XM},$$

$$h(\rho) = \sqrt{\frac{1+|\rho_{XM}|}{1-|\rho_{XM}|}} \text{ for } P'_{XM},$$

$$h(\rho) = \sqrt{\frac{1-|\rho_{MY}|}{1+|\rho_{MY}|}} \text{ for } P_{MY},$$

$$h(\rho) = \sqrt{\frac{1+|\rho_{MY}|}{1-|\rho_{MY}|}} \text{ for } P'_{MY},$$

$$h(\rho) = \sqrt{\frac{1-\text{sign}(\rho_{XM}\rho_{MY})\rho_{XY}}{1+\text{sign}(\rho_{XM}\rho_{MY})\rho_{XY}}} \text{ for } P_{XY},$$

$$h(\rho) = \sqrt{\frac{1+\text{sign}(\rho_{XM}\rho_{MY})\rho_{XY}}{1-\text{sign}(\rho_{XM}\rho_{MY})\rho_{XY}}} \text{ for } P'_{XY}.$$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-023-02298-9.

Acknowledgements This project came about initially as a class project by P. Bogdan in M. Regenwetter's course on "Modeling Heterogeneity." Regenwetter coordinated the ensuing three-person collaboration. P. Bogdan and M. Regenwetter roughly equally shared most of the writing of the main manuscript, with V. Cervantes providing the technical results and some writing support. V. Cervantes contributed nearly all mathematical findings (with P. Bogdan independently simulating most results on a computer), contributed nearly all of the writing for the Appendix, and developed nearly every aspect of our open-access shinyapp.

P. Bogdan carried out much of this work while supported by a *Thomas and Margaret Huang Graduate Fellowship* provided by the Beckman Institute for Advanced Science and Technology and a *Dissertation Completion Fellowship* provided by the University of Illinois. V. Cervantes carried out most of this work as an *Illinois Distinguished Postdoctoral Researcher* at the University of Illinois at Urbana-Champaign. The authors are not aware of any conflicts of interest.

Regenwetter has previously presented this work in colloquium talks at the Universität zu Köln, the University of Illinois at Urbana-Champaign, and Uppsala University.

The authors are grateful to Aaron Benjamin, Aron Barbey, Meichai Chen, Brittney Currie, Klaus Fiedler, Daniel Heck, Emily Neu Line, Julia Radu, Maria Robinson, David Trafimow, Haley V. West, as well as the colloquium audiences at the Universität zu Köln and the University of Illinois at Urbana-Champaign, and the workshop attendees at Uppsala University for feedback on earlier drafts and presentations. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the University of Illinois.

References

Allport, G. W. (1937). Personality: A psychological interpretation. *Holt*.

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review*, 37–47.

Araújo, M., Quintino, M. T., Budroni, C., Cunha, M. T., & Cabello, A. (2013). All noncontextuality inequalities for the n-cycle scenario. *Physical Review A, 88*(2), 022118. https://doi.org/10.1103/PhysRevA.88.022118

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychological Methods, 11*(2), 142.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81*(6), 506.

Bergman, L. R., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology, 9*(2), 291–319.

Bogat, G. A., von Eye, A., & Bergman, L. R. (2016). Person-oriented approaches. In: D. Cicchetti (Ed.) Developmental Psychopathology, 1–49.

Carroll, K. M. (2021). The profound heterogeneity of substance use disorders: Implications for treatment development. *Current Directions in Psychological Science, 30*(4), 358–364.

Chen, M., Regenwetter, M., & Davis-Stober, C. (2020). Collective choice may tell nothing about anyone's individual preferences. *Decision Analysis*. https://doi.org/10.1287/deca.2020.0417

Chmura Kraemer, H., Kiernan, M., Essex, M., & Kupfer, D. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology, 27*(2S), S101–S108.

Clauser, J. F., Horne, M. A., Shimony, A., & Holt, R. A. (1969). Proposed experiment to test local hidden-variable theories. *Physical Review Letters, 23*, 880–884. https://doi.org/10.1103/PhysRevLett.23.880

Collins, L. M., Graham, J. J., & Flaherty, B. P. (1998). An alternative framework for defining mediation. *Multivariate Behavioral Research, 33*(2), 295–312.

Condorcet, M. (1785). Essai Sur L'application de L'analyse a la Probabilite des Decisions Rendues A la Pluralite des Voix (Essay on the Application of the Probabilistic Analysis of Majority Vote Decisions). Paris: Imprimerie Royale.

Courbalay, A., Deroche, T., Prigent, E., Chalabaev, A., & Amorim, M.-A. (2015). Big five personality traits contribute to prosocial responses to others' pain. *Personality and Individual Differences, 78*, 94–99.

Danner, D., Hagemann, D., & Fiedler, K. (2015). Mediation analysis with structural equation models: Combining theory, design, and statistics. *European Journal of Social Psychology, 45*(4), 460–481.

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review, 126*, 865–879.

Dzhafarov, E. N., & Kujala, J. V. (2016). Context-content systems of random variables: The Contextuality-by-Default theory. *Journal of Mathematical Psychology, 74*, 11–33. https://doi.org/10.1016/j.jmp.2016.04.010

Edwards, J. R., & Lambert, L. S. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1.

Erev, I., & Feigin, P. (2022). Heterogeneous heterogeneity: Comment on Regenwetter et al. (2022). *Decision*, 118–120.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12*, 403–408.

Fairchild, A. J., & MacKinnon, D. P. (2014). Using mediation and moderation analyses to enhance prevention research. *Defining Prevention Science*, 537–555.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests–an analysis of articles published in 2015. *Journal of Experimental Social Psychology, 75*, 95–102.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology, 47*(6), 1231–1236.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences, 115*(27), E6106–E6115.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). mvtnorm: Multivariate normal and t distributions [Computer software manual]. Retrieved from https://CRAN.Rproject.org/package=mvtnorm (R package version 1.1-1)

Graziano, W., Habashi, M., Sheese, B., & Tobin, R. (2007). Agreeableness, empathy, and helping: A person× situation perspective. *Journal of Personality and Social Psychology, 93*(4), 583–599.

Grice, J. W., Cohn, A., Ramsey, R. R., & Chaney, J. M. (2015). On muddled reasoning and mediation modeling. *Basic and Applied Social Psychology, 37*(4), 214–225.

Hayes, A. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research, 50*(1), 1–22.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Howe, G. W., Beach, S. R., Brody, G. H., & Wyman, P. A. (2016). Translating genetic research into preventive intervention: The baseline target moderated mediator design. *Frontiers in Psychology, 6*, 1911.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309.

James, L. R., & Brett, J. M. (1984). Mediators, moderators, and tests for mediation. *Journal of Applied Psychology, 69*(2), 307.

Karazsia, B. T., & Berlin, K. S. (2018). Can a mediator moderate? Considering the role of time and change in the mediator-moderator distinction. *Behavior Therapy, 49*(1), 12–20.

Kellen, D. (2022). Behavioral decision research is not a Linda problem: Comment on Regenwetter et al. (2022). *Decision*, 112-117.

Kellen, D., & Klauer, K. C. (2019). Theories of the Wason selection task: A critical assessment of boundaries and benchmarks. Computational Brain & Behavior, 1–13.

Kendall, M. G., & Stuart, A. (1958). *The advanced theory of statistics*. London: Griffin.

Konstantopoulos, S., Li, W., Miller, S., & van der Ploeg, A. (2019). Using quantile regression to estimate intervention effects beyond the mean. *Educational and Psychological Measurement, 79*(5), 883–910.

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23*(4), 418–444.

Lee, H., Cashin, A. G., Lamb, S. E., Hopewell, S., Vansteelandt, S., Vander- Weele, T. J., ..., & Henschke, N. (2021). A guideline for reporting mediation analyses of randomized trials and observational studies: The AGReMA statement. *Journal of the American Medical Association, 326*(11), 1045–1056.

MacKinnon, D. P. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice, 21*(6), 675–681.

MacKinnon, D. P. (2012). *Introduction to statistical mediation analysis*. Routledge.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*, 173–181.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83.

Magnusson, D., & Bergman, L. R. (1988). *Individual and variable-based approaches to longitudinal research on early risk factors*. Cambridge University Press.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods, 12*(1), 23.

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research, 46*(5), 816–841.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*(2), 361.

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science, 18*(2), 112–117.

Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology, 89*(6), 852.

Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods, 26*(2), 255.

Owen, D. B. (1980). A table of normal integrals. *Communications in Statistics-Simulation and Computation, 9*, 389–419. https://doi.org/10.1080/03610918008812164

Owens, M., Stevenson, J., Hadwin, J., & Norgate, R. (2012). Anxiety and depression in academic performance: An exploration of the mediating factors of worry and working memory. *School Psychology International, 33*(4), 433–449.

Pillow, D. R., Sandler, I. N., Braver, S. L., Wolchik, S. A., & Gersten, J. C. (1991). Theory-based screening for prevention: Focusing on mediating processes in children of divorce. *American Journal of Community Psychology, 19*, 809–836.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*(1), 185–227.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77–98.

Regenwetter, M., & Robinson, M. (2017). The construct-behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review, 124*, 533–550.

Regenwetter, M., & Robinson, M. (2019). The construct-behavior gap revisited: Reply to Hertwig and Pleskac (2018). *Psychological Review, 126*, 451–454.

Regenwetter, M., & Robinson, M. M. (2022). Reply to Commentaries: Why should we worry about scientific conjunction fallacies? *Decision*, 124– 130.

Regenwetter, M., Robinson, M. M., & Wang, C. (2022). Are you an exception to your favorite decision theory? Behavioral decision research is a Linda problem! *Decision, 9*, 91–111.

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*(1), 3.

Saylors, R., & Trafimow, D. (2020). Why the increasing use of complex causal models is a problem: On the danger sophisticated theoretical narratives pose to truth. *Organizational Research Methods*, 1–14.

Scheibehenne, B. (2022). Experimenter meets correlator: Comment on Regenwetter et al. (2022). *Decision*, 121–123.

Simpson, J., Collins, W., Tran, S., & Haydon, K. (2007). Attachment and the experience and expression of emotions in romantic relationships: A developmental perspective. *Journal of Personality and Social Psychology, 92*(2), 355–367.

Smyth, H. L., & MacKinnon, D. P. (2021). Statistical evaluation of personoriented mediation using configural frequency analysis. *Integrative Psychological and Behavioral Science, 55*, 593–636.

Sobel, M. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290–312.

Sterba, S. K., & Bauer, D. J. (2010). Matching method with theory in personoriented developmental psychopathology research. *Development and Psychopathology, 22*(2), 239–254.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthèse, 48*, 191–199.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology, 37*(4), 235–246.

Thoemmes, F., & Lemmer, G. (2019). Mediation analysis revisited again. *Australasian Marketing Journal, 27*(1), 52–56.

Thomas, J. (2013). Association of personal distress with burnout, compassion fatigue, and compassion satisfaction among clinical social workers. *Journal of Social Service Research, 39*(3), 365–379.

Thorndike, R. M., Cunningham, G. K., Thorndike, R. L., & Hagen, E. P. (2010). *Measurement and evaluation in psychology and education* (8th ed.). Macmillan Publishing Co, Inc.

Trafimow, D. (2017). The probability of simple versus complex causal models in causal analyses. *Behavior Research Methods, 49*(2), 739–746.

von Eye, A., & Bergman, L. R. (2003). Research strategies in developmental psychopathology: Dimensional identity and the person-oriented approach. *Development and Psychopathology, 15*(3), 553–580.

von Eye, A., Mun, E. Y., & Mair, P. (2009). What carries a mediation process? Configural analysis of mediation. *Integrative Psychological and Behavioral Science, 43*, 228–247.

von Eye, A., & Wiedermann, W. (2022). *Configural frequency analysis: Foundations, models, and applications*. Springer Nature.

Wiedermann, W., & von Eye, A. (2021). A simple configural approach for testing person-oriented mediation hypotheses. *Integrative Psychological and Behavioral Science, 55*, 637–664.

Wiedermann, W., Zhang, B., Reinke, W., Herman, K. C., & von Eye, A. (2022). Distributional causal effects: Beyond an 'averagarian' view of intervention effects. *Psychological Methods*.

Witkiewitz, K., van der Maas, H. L., Hufford, M. R., & Marlatt, G. A. (2007). Nonnormality and divergence in posttreatment alcohol use: Reexamining the project match data "another way". *Journal of Abnormal Psychology, 116*(2), 378.

Authors and Affiliations

Paul C. Bogdan[1] · Víctor H. Cervantes[2] · Michel Regenwetter[2,3,4]

Víctor H. Cervantes
victorhc@illinois.edu

Michel Regenwetter
regenwet@illinois.edu

[1] Center for Cognitive Neuroscience, Duke University, 308 Research Dr, Durham 27708, NC, USA

[2] Department of Psychology, University of Illinois at Urbana-Champaign, 603 E Daniel St, Champaign 61820, IL, USA

[3] Department of Political Science, University of Illinois at Urbana-Champaign, 1407 W Gregory Dr, Urbana 61801, IL, USA

[4] Department of Electrical & Computer Engineering, University of Illinois at Urbana-Champaign, 306 N Wright St, Urbana 61801, IL, USA