Synthese (2021) 198:6689–6715
https://doi.org/10.1007/s11229-019-02484-3

Heritability, causal influence and locality

Pierrick Bourrat[1,2]

Received: 9 November 2018 / Accepted: 20 November 2019 / Published online: 4 December 2019

Abstract
Heritability is routinely interpreted causally. Yet, what such an interpretation amounts to is often unclear. Here, I provide a causal interpretation of this concept in terms of range of causal influence, one of several causal dimensions proposed within the interventionist account of causation. An information-theoretic measure of range of causal influence has recently been put forward in the literature. Starting from this formalization and relying upon Woodward's analysis, I show that an important problem associated with interpreting heritability causally, namely the locality problem, amounts, at least partly, to a low invariance and low stability between the genotype/environment and the phenotype of individuals. In light of this, I plead for a causal interpretation of heritability that takes the notions of Woodward's invariance and stability into consideration. In doing so, I defuse naive causal interpretations of heritability.

Keywords Heritability · Causality · Causal specificity · Causal influence · Locality · Invariance · Stability

1 Introduction

When it comes to estimate whether variations in genes as opposed to variations in the environment are the main cause of the observed variations in a phenotype, the classical approach is to estimate whether this phenotype is heritable. Yet, heritability is a statistical notion, and although it is routinely interpreted causally by scientists and the public, this interpretation has been controversial for the last 40 years (Lewontin 1974; Taylor 2006; Lynch and Bourrat 2017; Sesardic 2005; Tabery 2014; Tal 2009; Taylor 2010). Many authors have claimed, sometimes using disparaging language, that

✉ Pierrick Bourrat
p.bourrat@gmail.com

1 Department of Philosophy, Macquarie University, North Ryde, NSW 2109, Australia

2 Department of Philosophy and Charles Perkins Centre, The University of Sydney, Camperdown, NSW 2006, Australia

the use of heritability estimates to approach causality is inappropriate especially in the context of human behavioral genetics (for reviews including the historical context of these claims see Sesardic 2005, pp. 23–27; Tabery 2014, Chap. 3; Taylor 2006). Others, while often recognizing the limits of such a use, have been less concerned about this causal interpretation (e.g., Sesardic 2005; Tal 2009). At the heart of heritability analyses and their interpretation(s) thus lies the question of causation.

The philosophy of science literature on causal explanations has overwhelmingly adopted the interventionist account of causation (Woodward 2003, 2010, 2013). Within this account, a variable C is regarded as a cause of a variable O, if an ideal intervention[1] changing the value C can produce a change in the value of O.[2] This is the minimal criterion of causation within this framework. An ideal intervention is defined as a change in the value of a variable—here C—that produces no other change at the time of the intervention (Woodward 2003, 2010). Given the prominence of the interventionist account of causation, it is worthwhile investigating the relationship between causality and the notion of heritability from the perspective of this account. It might not only provide a better way to understand in what sense genes or genotypes cause phenotypes when heritability is positive, but also allow comparison of causal explanations originating from heritability studies with causal explanations from other sciences. Having a clearer understanding of the links between causation and heritability is also important in the context where claims of genes 'causing' such and such trait are very commonly found in popular media. The interventionist account of causation has recently been used to investigate the problem of gene-environment covariance in heritability studies by Lynch and Bourrat (2017). However, Lynch and Bourrat did not provide a general account of heritability within this framework. This is the aim of this article, which can also be regarded as an extension of the analysis provided by Oftedal (2005).

One can distinguish at least two different types of philosophical issues associated with the topic of heritability. The first one concerns the concept(s) of heritability, what I will refer to as theoretical 'heritability', and how they should be understood. It aims at answering questions of the following type: "Given a formal characterization of heritability how should one interpret it?", "Is a causal interpretation of heritability warranted and if yes, what sort of causal interpretation, and under what conditions does this interpretation become restricted (and in what ways) or break down?" The second type of issue concerns the way(s) under which one might *estimate* or measure theoretical heritability. Because we do not have an a priori knowledge about the theoretical heritability, we need to find situations under which, following some assumptions, we can get some confidence that the measure we make corresponds to the theoretical value. This second issue is one of statistical inference from observed data or exper-

[1] *Ideal* interventions should be distinguished from *experimental* or *physical* interventions in the sense that they can correspond to changes that are not necessarily physically possible. For instance, an ideal intervention can be an intervention on the mass of the moon, while keeping the same gravitational force between the Earth and the Moon. It would be impossible in practice to do so, but that does not constitute a problem from the interventionist perspective. Woodward (2013), among others, provides some justifications for the idea that interventions need not be physically possible. By 'intervention' in this paper, I will mean 'ideal intervention'.

[2] I use 'O' instead of 'E' for characterising a generic effect variable to avoid the confusion with the variable environment that will be used later on.

iments. There are many ways to estimate heritability and some problems associated with the fact that different estimates make different assumptions and sometimes do not correspond to the same concept. Thus the two issues are intertwined. Yet, in this paper, following the same strategy as in Taylor (2012), I will be mostly concerned with the first type of issue.

I use formal tools developed within the interventionist account of causation to delineate the scope and limits of the causal interpretation of theoretical heritability. Although focusing on this issue might look at first more remote from the practice of scientists using at least one of the available heritability concepts, I follow Lynch and Walsh (1998, p. 171) in their view when referring to one such concept, namely narrow heritability (h^2), that: "[p]roviding an explicit definition eliminates the ambiguity of the usage of h^2 in theoretical contexts, but highlights the practical problems of estimation." The analysis provided here is made in the spirit of Lynch and Walsh's remark.[3]

Concretely, this means that I will not be concerned with the way the theoretical measures presented in this paper might be estimated in a real study. That said, the toy example I present will aim at being more realistic than the typical examples used in the philosophical literature, such as the 'redhead example,' in which the notion of gene-environment covariance and its relation with heritability is illustrated by various thought experiments in which redheaded or blue-eyed children are subjected to different abuses (see Sesardic 2005, pp. 90–93). It should be furthermore noted that although obtaining accurate estimates of heritability of human behavioral phenotypes is a very challenging task, especially regarding the traits studied in behavioral genetics, behavioral genetics is only one of the many contexts in which the notion of heritability is used. Many of these contexts do not have the same limits for estimating heritability more accurately because experiments are possible.

The outcome of my analysis is a sophisticated causal interpretation of heritability which is in stark contrast with some naive interpretations. By 'naive causal interpretation', I mean a causal interpretation in which genes are regarded as causing a phenotype without considering the population context in which this interpretation is made. I show that although the worries coming from the opponents of a causal interpretation are justified, the limits they identify are not specific to heritability, a point previously made in passing by Tal (2009, p. 91). They pertain to a more general problem with the notion of causation as it is used by scientists. In fact, applying their arguments with parity would lead to the conclusion that most relationships considered as causal in a wide range of special sciences are after all not causal. I argue that interpreting heritability causally requires some precautions, which the formal version of the interventionist account recently outlined makes explicit (see Korb et al. 2011; Pocheville et al. 2017; Griffiths et al. 2015).

The paper will run as follows. In Sect. 2, I present the notion of range of causal influence, a particular kind of causal specificity, as a way to distinguish causes that have the highest degree of control over an effect variable, from those that do not (Griffiths et al. 2015). Within this approach, it has been proposed that one needs a notion like

[3] More generally, postulating theoretical parameters and then attempting to estimate them from the available data is a general problem that the branch of statistics known as "parameter estimation" attempts to solve.

range of causal influence, where causes with a larger range of causal influence enable one to better manipulate their effects (Woodward 2010). In Sect. 3, I recall the classical definition of heritability, and in Sect. 4, I show that heritability is related, when using my causal interpretation, to one particular measure of range of causal influence involving the distinction between *actual* and *potential* difference makers after Waters (2007). In Sect. 5, I discuss what is known as the problem of locality (Sesardic 2005, pp. 75–80): heritability estimates obtained in one population cannot be extrapolated to other populations. Ever since Lewontin (1974), the problem of locality, among other problems,[4] has been one important reason to argue against a causal interpretation of heritability. I demonstrate that locality results from particular features of causal relationships in some heritability studies, namely, different distributions for the causal variable, a low causal invariance, and/or a low stability. From there, I argue that if one were to regard locality as an important obstacle for interpreting heritability causally, one would have to, by the same token, dismiss many causal claims in the special sciences. I plead for a view in which causal claims are contextualized so that they may be well-interpreted.

2 Range of causal influence

Suppose one wants to answer the question: "why do some humans develop skin cancer?". One proposes two possible explanations citing two different putative causes: either individuals have had a prolonged exposure to solar ultraviolet (UV) radiation, and/or they have a particular genotype prone to skin cancer. Both explanations might appear equally relevant for answering the question. In fact, it has been shown that both intermittent solar UV exposure, particularly in childhood (Gandini et al. 2005a) and genetic factors (Hayward 2003) are involved in some forms of skin cancers such as melanoma, the type of skin cancer with the highest morbidity.[5] Were it scientifically established that one of the two causes is more important for explaining the development of skin cancers, the interventionist minimal criterion of causation presented in the Introduction, by merely stating that both are causes, would not permit to choose between them. This is because there exist interventions on both variables that would lead some individuals to develop skin cancer or prevent them from developing it.

Considering this, any measure of the difference in importance or influence for each cause involved in the prevalence of skin cancer will be useful.[6] Recent formal approaches to causation have married the causal modeling approach (see Pearl 2009) with information theory to produce such a quantitative assessment of causal influence (Korb et al. 2011; Griffiths et al. 2015; Pocheville et al. 2017). Here I present in broad strokes Griffiths et al.'s account and extend it for cases in which information theory is not best suited.

[4] For other problems see for instance Tabery (2014), Sarkar (1998) and Northcott (2006, 2008).

[5] For instance some mutations on the genes CDKN2A and CDK4 have both been associated with melanoma (Begg et al. 2005; Hayward 2003) and some families are more prone to develop skin cancers than others (Gandini et al. 2005b).

[6] The analysis of variance is classically regarded as providing precisely this, but its causal interpretation has been contested (e.g. Northcott 2006).

They start from the notion of causal specificity. Causal specificity is an ambiguous notion since it can refer to at least two distinct dimensions of causal relationships identified by Woodward (2010). The first dimension refers to the extent to which, on average, considering a relationship that satisfies the interventionist minimal criterion of causation, one intervention on the causal variable corresponds to a precise change in the effect variable. In other words it refers to the extent to which a mapping between causal values and effect values is bijective, namely that to one value of the causal variable corresponds exactly one value of the effect variable. Woodward (2010, p. 310) refers to this as the "one cause-one effect" notion of causal specificity. I will refer to it as "one-to-one specificity". The second dimension refers to the range of causal influence a variable exerts on another variable. As first showed by Woodward (2003, pp. 218–220), this notion of causal specificity is closely related to Lewis' (2000) notion of influence (see also Weber 2006 and Waters 2007 for a similar use). The basic idea underlying this dimension of causal specificity is that the more influential a causal variable C is for a putative effect O, the higher the number of possible interventions on C that lead to changes in the values of O that no other intervention on C would lead to. I will refer to this second dimension, to which Griffiths et al. give a precise measure, as "range of causal influence" to avoid any ambiguity.[7]

Given a particular relationship $C \rightarrow O$ satisfying the interventionist minimal criterion, the probability distribution for C and the conditional probabilities of O for each value of C, one can use information theory to measure the range of causal influence of C on O. Griffiths et al argue that the right information theoretic measure to approach the range of influence is the amount of mutual information that interventions on C carry about O. I refer to this measure as "mutual *causal* information."[8] which is calculated as follows:

$$I(O; \widehat{C}) = H(O) - H(O \mid \widehat{C}), \tag{1}$$

where $I(O; \widehat{C})$ is the mutual causal information from C to O, $H(O)$ is the entropy of O, that is the amount of uncertainty on O, and $H(O \mid \widehat{C})$ is the entropy of O knowing the value set for C (the hat on a C represents the fact that its values are set by ideal interventions), that is the amount of uncertainty remaining on O when we already know the value set for C. For more details on this measure and the notions of entropy and conditional entropy see Griffiths et al. (2015); for a longer yet accessible introduction to information theory see Stone (2015).

Information theory is best suited for nominal variables, that is variables with which no form of mathematical computation can be performed (Stevens 1946). Yet, in biology variables are generally quantitative, for which the analysis of variance provides quantities corresponding to entropy and mutual information (Garner and McGill 1956). A version of range of causal influence for quantitative variables reads:

$$V(O : \widehat{C}) = V(O) - V(O \mid \widehat{C}), \tag{2}$$

[7] For in-depth treatments of these two notions of causal specificity see Bourrat (2019a, b).

[8] For more details on the properties of this measure see Griffiths et al. (2015) and Pocheville et al. (2017).

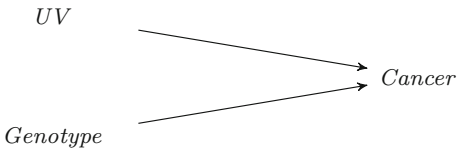


Fig. 1 A causal diagram representing the causal dependencies between variables involved in the skin cancer example (error terms are ignored). The variable *Genotype* represents the genotype of an individual; the variable *UV* measures the monthly highest UV levels on the global UV index in the population; the variable *Cancer* measures the probability for an individual to develop a skin cancer during a fixed period of time

where $V(O : \widehat{C})$ represents the amount of variance in O explained by the variance in C, with values of C being set by interventions, $V(O)$ is the variance of O, and $V(O \mid \widehat{C})$ is the variance of O knowing the value set for C.[9] Note, and this is important, that Eq. (2) encompasses—in the sense that they are inseparable—both a measure of the range of causal influence and a measure of what one might call the "causal strength" of the relationship. Causal strength measures the extent to which intervening on C by one unit, produces a change in O by x units, where 'x' measures the strength of the association. One difference between Eqs. (1) and (2), is that the measure obtained in the former is in bits while the latter is in the unit of the effect variable (more on the difference between the two equations below).

Take the skin cancer example. To operationalize the range of causal influence measure [whether using Eq. (1) or Eq. (2)], one first needs to specify the causal dependencies at stake. These dependencies can be represented on a causal diagram (see Fig. 1, since I am not concerned here with statistical inference, error terms are ignored). Let us suppose that each value of the variables (either nominal or quantitative) *Genotype* and *UV* can be given a probability distribution (or probability density function in the case of quantitative continuous variables) and that we know the conditional probabilities of the variables *Cancer* for each value of *Genotype* and *UV*. With this in place, one can then calculate the range of causal influence between the two cause-effect pairs of variables using Griffiths et al's measure (Eq. 1) or its corresponding measure using variances (Eq. 2). The ranges of influence obtained for *Genotype* $\rightarrow$ *Cancer* and *UV* $\rightarrow$ *Cancer* can then be compared to decide which one of the two causal variables has a greater range of causal influence.

To see this in more detail, let us assume that the variable *Genotype* has only four possible values with equal probability (0.25), namely $\{G_1, G_2, G_3, G_4\}$. Based on the global UV index scale (World Health Organization 2002), to which I have removed the 'extreme' value to make the example simpler, we assume that the variable *UV* has also four values with equal probabilities (0.25) {*Low*, *Moderate*, *High*, *Very High*}, each corresponding to a risk of harm from unprotected sun exposure. Finally based on a realistic worldwide incidence rate of melanoma (see Stewart and Wild 2014, p. 496), we assume that the variable *Cancer* has also four equiprobable values (0.25), namely $\{1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4}, 4 \times 10^{-4}\}$.[10]

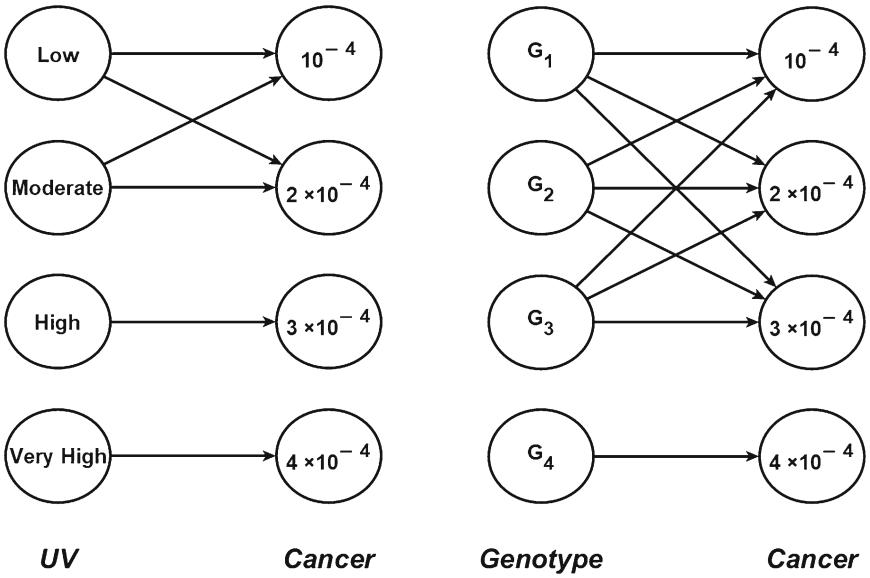
[9] I thank Arnaud Pocheville for proposing this measure.

[10] This example uses some nominal variables and will treat quantitative variables as if they were nominal. Illustrating the notion of range of causal influence is handier with nominal rather than quantitative variables, since, as mentioned above, Eq. (2) provides a combined measure of range of influence and causal strength.

To be able to calculate the range of causal influence, given that grain of description,[11] on the one hand, of *UV* on *Cancer*, and on the other hand, of *Genotype* on *Cancer*, we need to know the probability for each value of *Genotype* and *UV* to be associated with each value of *Cancer*. If we adopt the probabilities presented in the diagrams (a) and (b) of Fig. 2 (see the caption to understand how these probabilities are calculated) between the two relationships of the causal diagram presented in Fig. 1, we can see that intervening on the variable *UV* carries more mutual causal information on the variable *Cancer* than does the variable *Genotype*. Indeed, we find a mutual causal information of about 0.81 *bits* from *Genotype* to *Cancer* and 1 *bit* from *UV* to *Cancer*.[12] Concretely, this means that intervening on the value of *Genotype* has a lower probability to change the value of *Cancer* than intervening on the value of *UV* does. One can get an intuitive understanding of this by noticing that there are less values to choose from with *Genotype* than with *UV* and that there are, on average, more arrows leaving the causal variable *Genotype* than the causal variable *UV*. Had *Genotype*, *UV* and *Cancer* all been quantitative variables using Eq. (2) and a similar reasoning, we would find that the amount of variance of *Cancer* explained by intervening on *UV* is higher than when intervening on *Genotype*. This last remark comes however with one caveat, namely that the difference between two adjacent values for the causal effect variables should be the same across the range of possible values for these quantitative variables, different interventions should produce proportionate effects in a given direction (the causal relationship between the variables should be linear), and the causal strength of *Genotype* and *UV* on *Cancer* should be the same. This comes from the fact that with quantitative variables, that is variables of which the values on which some mathematical operations can be made, the dimension of causal strength is at play. This dimension does not appear with nominal variables because all one can say about the values of such variables is that they are different (one can only say that the color 'red' is different from the color 'green'; there is not a sense in which green is higher or lower than red). One consequence of this is that changing the values toward which the arrows point in the diagrams (a) and (b) of Fig. 2, will not change the value for range of causal influence obtained with Eq. (1) (whether the variables are nominal or quantitative). By assuming that the causal strength of the relationships *Genotype* $\rightarrow$ *Cancer* and *UV* $\rightarrow$ *Cancer* are linear and of the same strength, across the whole range of possible values for causes, means that, under these assumptions, causal strength is not a factor to take into consideration when comparing these relationships since it is the same for all the intervals of the variables considered.

[11] Note that there is in principle an infinity of ways in which a variable can be discretised, depending on the grain of description one uses. Starting from Yablo's (1992, p. 4) example of a pigeon pecking because she was exposed to a red stimulus or a scarlet stimulus, Pocheville et al. (2017) provide a criterion to choose the grain of description that is appropriate for providing a proportional explanation, by discretising the variable involved in the relationship. Proportionality is a dimension to take into consideration for providing adequate causal explanations as emphasized by Woodward (2010). Pocheville et al call their criterion 'proportionality constraint' and define it as follows: "Given an effect variable *O* that is a target of intervention or causal explanation, a causal variable *C* should be discretised so as to minimise the entropy of *C* whilst maximising specificity for *O*" (p. 272, 'E' has been replaced by 'O'). I will assume throughout the manuscript that the grain of description satisfying the proportionality constraint has been chosen.

[12] For a short tutorial on how to calculate the mutual information between two variables see Griffiths et al. (2015).



(a) Causal diagram of the range of causal influence of the variable UV on the variable $Cancer$.

(b) Causal diagram of the range of causal influence of the variable $Genotype$ on the variable $Cancer$.

Fig. 2 Causal diagram representing the causal relationships between *Genotype*/*UV* and *Cancer* (errors terms are ignored). In all the diagrams presented in this figure, Figs. 3 and 4, equiprobability for the values of C is supposed. Each arrow leaving from one value of C represents the probability of this value causing the value of O it points to. If only one arrow leaves a given value of *UV* or *Genotype*, the conditional probability on this value is 1 (e.g., for the arrow leaving the value "high" of *UV*). If more than one arrow leaves a value of *UV* or *Genotype*, then we suppose that each arrow has the same probability conditioning on this value (e.g., 0.5 for arrows leaving the value "low" of *UV* connecting to two values of *Cancer*, and $\frac{1}{3}$ for the arrows leaving the value G_1 connecting to three values of *Cancer*)

With the notion of range of causal influence now presented, in the next three sections, I turn to the relation between this notion and heritability. After having briefly presented the classical definition of heritability in the next section, I show in Sect. 4 how heritability relates to the notion of range of causal influence. In Sect. 5, in light of the interventionist approach developed in the preceding sections, I discuss one famous problem within the literature on heritability, namely the problem of locality. More particularly, to treat this problem, I appeal to the notions of causal invariance and stability developed within the interventionist account, which I briefly present.

3 Heritability

The notion of heritability originated at a time where DNA had not yet been discovered to be the material substrate for genetic information.[13] In consequence, the statistical techniques underlying its estimation do not rely on any measure of genetic or environmental physical factors, so that even the words 'gene' and 'environment' can be dropped from the analyses (Taylor 2012). Yet, in the Introduction I referred to estimating heritability as a way to elucidate whether genes or the environment are causes of phenotypic change, which is how it is now often interpreted. This difference might lead the reader to question the validity of an approach to heritability grounded within the interventionist account of causation which precisely supposes intervening on such factors when it is supposed that no such factors can be intervened upon in the classical methods. These two views on heritability might even be regarded as incommensurable paradigms.[14]

I have two things to respond to this. First, as mentioned in Footnote 1, there is in principle no obstacle to applying the interventionist account on variables that could not physically be intervened upon. Second, in the last twenty years, a lot of effort has been carried out to create a bridge between these two paradigms with the development and use of molecular biology methods to physically 'locate' genetic information and develop statistical methods in consequence, to get new estimates of heritability, such as genome-wide association studies (GWAS) from which single-nucleotide polymorphisms (SNPs) can be extracted (see Visscher et al. 2008; Bourrat and Lu 2017; Yang et al. 2010).[15] Although, as emphasized by a reviewer, a direct translation between traditional approaches to heritability and those relying on GWAS might not be as easy to make as believed by some such as Visscher and Goddard (2019), I will assume that such a link can be drawn.

This remark to the side, there exist several definitions of theoretical heritability and several ways to estimate it (see Falconer and Mackay 1996; Lynch and Walsh 1998; Bourrat 2015; Downes 2009; Godfrey-Smith 2009; Jacquard 1983; Sarkar 1998; Taylor 2012; Tal 2009). To avoid confusion, following the footstep of others (see Lynch and Walsh 1998; Taylor 2006, 2010) it is important to be clear about which concept I will refer to here. A useful starting point in this literature, is the definition of heritability as the proportion of phenotypic variance that can be attributed to genetic variance (Falconer and Mackay 1996, p. 160). This definition follows a linear model in which the phenotype of an individual (P) is the dependent variable of one independent variable, namely the genotype of this individual (G), and a combined deviation due to the environment, genotype-environment interaction and noise (E),[16] of which the mean is 0:

[13] See for instance Fisher (1918) and (Wright 1920) for early uses of the concept of heritability. Note that the origin of the term 'heritability' itself is debated (see Bell 1977).

[14] This point has been raised by an anonymous reviewer.

[15] For more on taking this approach with heritability and how to frame it within the interventionist account, see Bourrat (accepted).

[16] In a situation where it is assumed there is no genotype-environment interaction and no noise, E is the deviation due to the environment.

$$P = G + E \tag{3}$$

From Eq. (3) and the further assumptions that there is no correlation and no interaction between G and E, we can deduce, following the properties of variances, that the phenotypic variance in a population is equal to the sum of the genotypic variance $V(G)$ and environmental variance $V(E)$:[17]

$$V(P) = V(G + E) = V(G) + V(E) \tag{4}$$

The (broad sense) heritability of P (H^2) is then defined as the ratio of $V(G)$ on $V(P)$:

$$H^2 = \frac{V(G)}{V(P)} \tag{5}$$

The variance of a variable is a statistical measure of the dispersion of that variable's values in a population. Heritability is consequently a statistical measure, but it is often interpreted causally as the level of causal influence of the genotype on the phenotype. The move from a statistical description to a causal interpretation is quite straightforward in the linear model proposed above in which G and E make independent contributions. Using the casual diagram in Fig. 1, one dimension of causation corresponds to the relative weight of the arrow going from *Genotype* to *Cancer*, when compared to the arrow going from *UV* to *Cancer*. There are important problems associated with a causal interpretation when the assumptions of no gene-environment correlation and interaction are violated (Lewontin 1974; Lynch and Bourrat 2017; Sesardic 2005; Tal 2009, 2012; Taylor 2006, 2010; Moffitt et al. 2005). Tal (2012) provides a useful method to account for the extent to which this interpretation can be given in the contexts where gene-environment covariance and interaction are important factors. In this section and most of the paper, I consider the simplest linear case. I will come back to the issue of linearity in Sect. 5.

Note also that I assume here that G corresponds to *Genotype*, E corresponds to *UV*, and P corresponds to *Cancer*. Such an interpretation is not benign. First, it implies a concept of genotype used as one corresponding to a particular material substrate, here the mutations associated with the skin cancer. This is however not the notion of the gene classically used in quantitative genetics, which rather corresponds to an informational notion of the gene which has no particular material substrate. The distinction between these two conceptions of the gene, namely the informational and the molecular notion, as well as the possible tensions resulting from an ambiguous use of the terms 'gene' or 'genetic' has been pointed out numerous times in the literature (e.g., Taylor 2009, 2010, 2012; Griffiths and Neumann-Held 1999; Griffiths and Stotz 2013; Bourrat and Lu 2017; Lu and Bourrat 2018). Second, for similar reasons, interpreting E as *UV* implies that there would be no other directional physical factors in the environment varying between genotypes that would be involved in the development of skin cancer.

[17] The terms 'genotypic variance' and 'environmental variance', though canonical, can be confusing. The reader should understand them as 'variance in phenotype attributed to genotype variation' and 'to environmental variation', respectively (see Taylor 2006, 2010, 2012, for discussions of the sense in which these terms can be confusing).

This should, of course, only be regarded as an idealization (which I will relax later on). In the particular examples given in Fig. 2, because the relationship between the causal relata are indeterministic (to one causal value can correspond several effect values) some part of the variation in E comes from the background which can be regarded as noise. Nevertheless considering the opening remark of this section these assumptions do not pose any particular problem here, so long as one is aware that direct correspondences between the informational and molecular notions of the gene are anything but necessary.

It should also be noted that, considering heritability relies on the statistical analysis known as the analysis of variance (ANOVA), the notion of causality it rests upon refers to populations rather than to the individuals forming populations (Northcott 2006, 2008). This means that if one were to obtain a heritability of 1, considering the above assumptions are fulfilled, one could not interpret this value as meaning that genes, and not the environment, cause a given individual to have a skin cancer susceptibility. Rather, a correct interpretation would be that *in a population*, intervening on the genotype of individuals will have an effect on the susceptibility of skin cancer of these individuals, while changing the environment (UV) will not.[18] I will come back to this point in the next section when I discuss the two types of difference makers drawn within the interventionist account.

As mentioned above, there are other senses and definitions of heritability (Jacquard 1983; Lynch and Walsh 1998; Falconer and Mackay 1996; Downes 2009; Tal 2009), such as narrow-sense heritability, and the ways to estimate different heritabilities also vary from one discipline to another (see Visscher et al. 2008). My analysis can straightforwardly be extended to these other senses and estimates. For that reason, I want to stress that I intend this analysis to serve any discipline using a concept of heritability rather than solely behavioral genetics, a discipline in which some naive causal interpretations of heritability have led to not only erroneous claims on human differences, but also potentially harmful ones. This is the primary reason why the main example I use throughout is not one stemming from behavioral genetics.

4 Range of causal influence and heritability

We saw in the previous section that, in simple additive cases, heritability can and has been given a straightforward causal interpretation. In this section, I make the link between heritability and causation tighter by showing that the notion of range of causal influence presented in Sect. 2 captures at least partly the intuition underlying the causal interpretation of heritability. I start by presenting some similarities between range of causal influence and heritability and then propose to formalize them. Important to my discussion will be the distinction between actual and potential difference makers, after Waters (2007).

[18] See however Tal (2009), who provides a probabilistic individual interpretation from heritability measures under some simple assumptions as well as a discussion of the tension between an absolute individual causal interpretation of heritability and an interpretation contextualized within a population in which there is variation.

When discussing heritability and causation, it is commonplace to start with the remark that all phenotypes are the result of a causal interaction between a genotype and its environment and therefore that both genes and the environment are causes for a phenotype. This view is known as the interactionist consensus (Oftedal 2005; Sesardic 2005, pp. 48–56; Sterelny and Griffiths 1999, pp. 97–100). Gene-environment interaction, in this sense, implies that the claim from scientists that differences in genes are causally associated with differences in phenotypes is not equivalent to the claim that the environment is not causally involved in the production of phenotypes. Similarly, when the same scientists claim that differences in genes are not causally associated with differences in the phenotypes, they do not mean that genes are not causally involved in the production of the phenotype. This relates to the above point that the notion of causation invoked with regards to heritability is one that refers to a *population* rather than to a more intuitive notion of causation referring to the *individuals* forming this population (Northcott 2006, 2008).[19] Concretely thus, the claim that differences in genes *are (not) causally associated* with differences in phenotypes should not be understood as genes being necessary (unnecessary) for the production of a phenotype. One can already notice some similarity between this understanding of causation and the notion of cause used within the interventionist account. Recall that from the interventionist minimal criterion for causation, the only requirement for C to be a cause of O is that there exists an intervention changing the value of C that produces a change in the value of O, not that an absence of C produces an absence of O, even though in some particular situations the presence or absence of a cause will be the relevant values.

With this distinction made, one will be tempted to make two claims about heritability once interpreted causally. The first one is that if there is no genotypic variation in a population, that is there is only one value for G, heritability is nil since $V(G)$ is, by definition, nil. This claim is correct. Without genotypic variation, genotypic variance, which is a statistical descriptor of variation, will necessarily be nil and so will heritability.

The second tempting claim to make is that if genotypic variation exists in the population, it necessarily means that heritability is positive. This second claim is, however, false: Genotypic variation is not necessarily associated with phenotypic variation. The reason why this second claim is false can be illustrated with an example. Suppose a population in which there is some genotypic variation, so that there are individuals with different combinations of alleles, understood here as sequences of DNA with specific nucleotidic combinations—see the above point about the different notions of the gene and why this is only an idealization—and yet in which $V(G)$ and consequently heritability are both nil. This type of situation might look at first paradoxical. But the paradox quickly dissolves if one remembers that genotypic and environmental variance, in heritability studies, are defined in units of phenotypes, not in units of genotypes and environment respectively. Recall that G and E are *components of* P and therefore have the same unit as P. From this point, Lewontin (1974, pp. 402–403)

[19] "The distinction between individual and population cause is a tricky one. I use it here in a very intuitive way. For a deeper analysis, see Bourrat (2019)" Bourrat, Pierrick. 'Evolution Is about Populations, but Its Causes Are about Individuals'. Biological Theory, 12 October 2019. https://doi.org/10.1007/s13752-019-00329-3

concludes, when referring to $V(G)$ and $V(E)$, that "[w]e are not actually assessing how much variation in environment or genotype exists, but only how much *perturbation* of phenotype has been the outcome of average difference in environment [or genotype]" (emphasis added). Put in interventionist terms, Lewontin's remark implies that the variances do not measure whether there is some variation on G or E, but rather to what extent the intervening on G and E leads to change in P (the higher the value, the higher the range of interventions and magnitude of effect). For instance, if an intervention on genotype produces no change in phenotype, this will not be captured by $V(G)$.[20]

This feature of the heritability statistics is strikingly similar to one feature of range of causal influence. As shown by Pocheville et al. (2017), the range of causal influence between two variables—which they refer to as 'causal specificity—when measured as mutual causal information, is insensitive to differences in the values of a causal variable that do not lead to differences in the values of the effect variables. That is to say that C in the diagram (a) of Fig. 3 has the same range of causal influence on O as C on O in the diagram (b) of the same figure.[21] In the same way as different genotypes associated with the same phenotype have no impact on the value of heritability, different values of C associated with the same values of O do not change the range of causal influence. As I have argued elsewhere (see Bourrat 2019b), the difference between the diagrams of Fig. 4 corresponds to a difference in one-to-one specificity, with the causal relationship in diagram (b) being more one-to-one specific than the relationship in diagram (a). This type of causal specificity is important in some contexts. Heritability measures, like measures of range of causal influence, do not permit, however, to capture these differences.

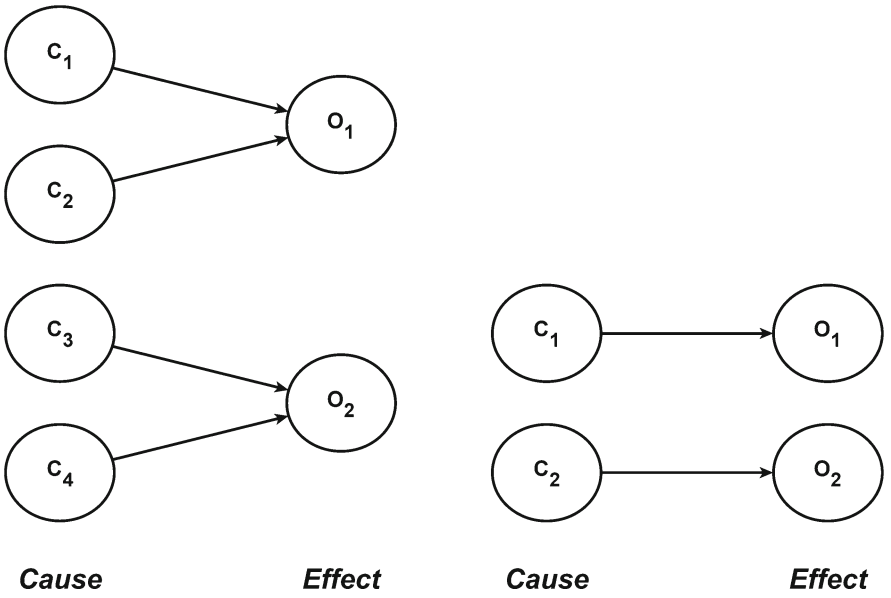
Before turning to another similarity between heritability and range of causal influence, a further distinction used in the interventionist literature must be made, namely the distinction between potential and actual difference makers (Waters 2007). I will come back to this distinction when I discuss the problem of locality in the next section. A potential difference maker with respect to O is a variable C on which at least one *possible* intervention from one value to another of C would produce a change in the probability distribution of O. An actual difference maker with respect to O is a variable C in a given population[22] which is a difference maker on which at least one intervention from one value to another of C produces a change in the value of O which is *actually observed* in the population.

An actual difference maker in one population might not be one in another population (or it might be to a greater or lesser extent). This is for two types of reasons. First, the background in one population might be different from the background of another population and the respective backgrounds might interact differently with the relationship $C \rightarrow O$ in each population. In some extreme cases this difference might lead to changes in O in one population and no changes in the other. In this set of cases,

[20] Note that in cases where variables are nominal, for which information theory is best suited, the units of phenotype would be in bits. By definition, nominal variables have no metrics.

[21] It is from this point that Pocheville et al. (2017) devise their 'proportionality constraint' for choosing the right grain of description for a causal relationship (see foonote 11).

[22] The notion of population corresponds here to a population of events, not necessarily a population of individuals, although in the case of heritability it will be a population of individuals.



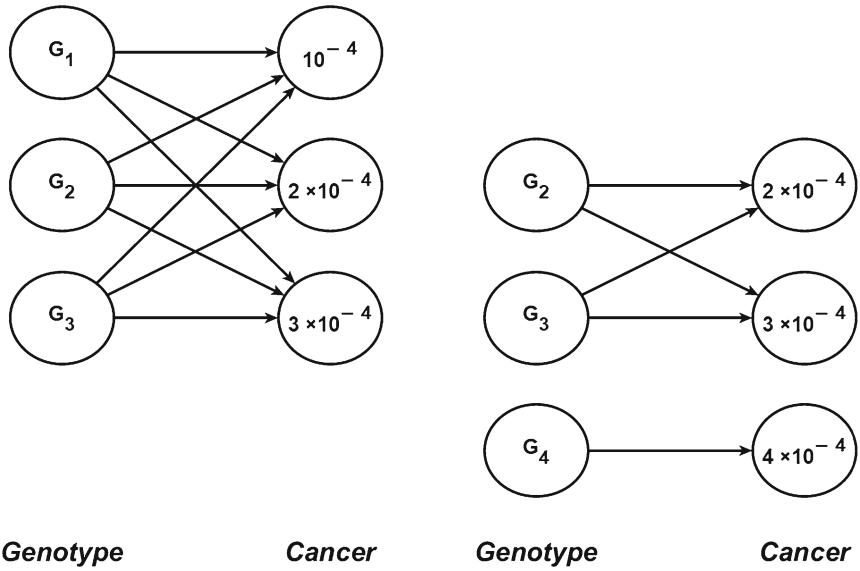
(a) Causal diagram of the relationship $C \rightarrow O$ with four C values and two O values. Range of causal influence is 1 *bit*.

(b) Causal diagram of the relationship $C \rightarrow O$ with two C values and two O values. Range of causal influence is 1 *bit*.

Fig. 3 Causal diagrams presenting two causal relationships with the same range of causal influence in spite of a different number of values for C

C will be an actual difference maker in the former population, but not in the latter. Second, assuming a constant background, some values of C that are causally associated with some discriminate values of O might be more or less frequent in different populations. In some other extreme cases the frequency of values of C associated with discriminate values of O might be 0 in one population and > 0 in another. In this set of cases, C will not be an actual difference maker in the former population, but will be one in the latter population.

Interpreting the lessons from the interactionist consensus using the actual/potential difference maker distinction is quite straightforward. Indeed, since an intervention on G or E could *potentially* be the value 'absent,' both of which would produce a value of 'no phenotype' for P, both G and E are potential difference makers. Heritability is always estimated within a population (whether it is an actual population or an artificial one due to the experimental design) in which only subsets of the possible values of G and E exist. Consequently, if heritability is interpreted causally as a measure of the causal influence of the genotype on the phenotype in a population, the right concept to use here is that of *actual*, not potential difference maker. In fact, because populations vary in genotypic composition and in the environment in which they are found (more generally the background conditions), the relationships between genotypes and phenotype and between the environment and the phenotype



(a) Causal diagram of the relationship $Genotype \rightarrow Cancer$ in Pop_A. Range of causal influence is 0 bit.

(b) Causal diagram of the relationship $Genotype \rightarrow Cancer$ in Pop_B. Range of causal influence is 0.92 $bits$.

Fig. 4 Causal diagrams representing the causal relationships between *Genotype* and *Cancer* in two hypothetical populations

might be different in different populations. This means, for instance, that intervening on G might make no difference to P in a given population, in which case we would conclude that $V(G) = 0$ (and consequently $H^2 = 0$) in this population. Yet, the same intervention might make a difference to P in a different population, in which case we would conclude that $V(G) > 0$ (and consequently $H^2 > 0$) in this other population. The same reasoning applies for $V(E)$.

To see how this remark translates concretely into heritability studies, take again the example of skin cancer, and let us make some further assumptions. Suppose now that both having either of the three genotypes G_1, G_2 or G_3 is causally independent from the chance of developing a skin cancer, and having the genotype G_4 causes, independently from any other factors, skin cancer with some non-nil probability following the causal diagram presented in the diagram (b) of Fig. 2. We might assume that G_4 individuals are carriers of a mutation on the gene CDKN2A or the gene CDK4, which I mentioned earlier (see Footnote 5).

Suppose now that we want to know the heritability of skin cancer and that we launch a study in a population Pop_A in which only the genotypes G_1, G_2 and G_3 are present, each with the same frequency ($\frac{1}{3}$). In this population, some individuals with either of the three genotypes will develop a skin cancer, for instance because these individuals expose their skin to the sun more than others. Assume, however, no systematic differences in sunbathing habits between the three genotypes. Because there are no individuals with G_4, the relationship between *Genotype* and *Cancer* is

represented in the diagram (a) of Fig. 4. This diagram is a subset of the diagram (b) of Fig. 2. In this example, because the variation in phenotype resulting from the variation in the genotype is nil, a measure of heritability would consequently be nil.[23]

We can now ask what would the measure of causal influence obtained from Eq. (1)[24] be within the interventionist account in such a case.[25] Since intervening on the genotype of individuals in Pop_A, with the actual frequencies of this population changing from one value to another value (e.g., from G_1 to G_3), leads to no change in P, G is not an actual difference maker in this population. A measure of mutual causal information from Eq. (1) leads to the same conclusion since there are 0 $bits$ of mutual causal information from G to P, which means that G, in this population, carries no mutual information about the probability of developing a skin cancer.[26] Using this example, we can thus see that a nil heritability corresponds to a case of nil actual range of causal influence of G on P. What is true when considering this extremely simple example can be generalized to any case in which intervening on the genotype, using the probability distribution of the focal population, does not change the value of the phenotype.

Suppose now that we study the incidence of cancer in a second population (Pop_B). In this population there are no individuals with genotype G_1 but equal frequencies of individuals with genotypes G_2, G_3 and G_4. Here again we assume no difference in sunbathing habits between the three genotypes. In this population, individuals with genotype G_4 have a higher probability to develop a skin cancer due to the fact that they are carriers of an allele that increases their susceptibility to skin cancer. The relationship between *Genotype* and *Cancer* is represented in the diagram (b) of Fig. 4 which, in this case too, is merely a subset of the diagram (b) of Fig. 2.

In Pop_B contrary to Pop_A, because *actual* variation in phenotype resulting from a variation in the genotype is positive, we have a positive heritability. Since intervening on the genotype of individuals in Pop_B, with the actual frequencies of this population, leads to some change in *Cancer*, G is an actual different maker in this population. In fact, computing the mutual causal information shows that there are about 0.92 $bits$ of information from *Genotype* to *Cancer* which means that *Genotype*, in this population, and contrary to the case in Pop_A, causally influences the probability of getting a skin cancer.[27] Using this example, we can thus conclude that a positive actual range of influence of G on P would translate into a positive heritability. What is true with this example can be generalized to any case in which intervening on the genotype, using the probability distribution of the focal population, changes the value of the phenotype.

[23] We consider here that the variables *Genotype*, *UV* and *Cancer* correspond to the variables G, E and P, respectively, in heritability studies. This would imply that a link between genotypic variation and phenotype has been established as it is done in genome-wide association studies (see Visscher et al. 2008).

[24] We cannot use Eq. (2) here because the variable *Genotype* is nominal, and as mentioned earlier, it would not make sense in that context. Based on the assumptions of quantitative genetics (e.g.,following the infinitesimal model, assuming that an infinity of genes, with each of them having an infinitely small effect on phenotype) one can however interpret a set of nominal data as being continuous.

[25] This measure corresponds to the SAD (for "specific actual difference") of Griffiths et al. (2015).

[26] Similarly, had *Genotype*, *UV* and *Cancer* been quantitative variables, then using the measure from Eq. (2) would have led to the conclusion that G does not causally influence P.

[27] Had G been a quantitative variable, using the measure from Eq. (2) would have led to the conclusion that G causally influences P.

We have established thus far, with the skin cancer example, first that a nil actual causal influence of the genotype on the phenotype would correspond to a nil heritability, when interpreted causally within the interventionist account of causation; and second that some actual influence between a genotype and a phenotype would correspond to a positive heritability. Let us now see what distinguishes cases in which heritability is low, but not nil ($0 < H^2 < 0.5$) and cases in which it is high (> 0.5) in terms of range of causal influence.

Let us start with a case of low but not nil heritability ($0 < H^2 < 0.5$). For heritability to be low but not nil, following its definition given in Eq. (5), two conditions must be satisfied: a) genotypic variation (measured in units of phenotype) must be positive, so that genotypic variation results in a positive variation in P, but b) this variation is lower than the variation in phenotype that results from environmental variation.

A possible situation satisfying these two conditions can be given using our example of skin cancer. This situation implies that a change in probability to develop a skin cancer (P) depends much less, in the population, on whether the individuals have a given genotype than on the time they were exposed to the sun. Suppose a third population Pop_C in which the four genotypes (G_1, G_2, G_3 and G_4) are represented with the same frequency, that individuals have different sun exposure practices (but no differences, on average, between the four genotypes) and that we know the dose of UV received by each individual over a relevant period of time to predict the risk of skin cancer. For simplicity I suppose that individuals can be separated into four homogeneous groups for *UV* and that the relationship *UV* and *Cancer* is the same as the one proposed in the diagram (a) of Fig. 2.

Once these assumptions are made, we have variation in phenotype attributable to genotypic variation much smaller than variation in phenotype attributable to environmental variation. Since all the values of the causal variables in Pop_C have frequencies which are the same as the probabilities of each value of the example presented in Sect. 2, the computations we made about the example in that section will be the same here. These led to a smaller range of causal influence for the variable *Genotype* than for the variable *UV*, namely 0.81 $bits$ and 1 $bits$ respectively. From there it is tempting to conclude that a small heritability is at least partly equivalent to the actual range of causal influence of *UV* on *Cancer* being higher than that of *Genotype*. Thus, starting from the information theoretic measure for range of causal influence of Eq. (1), a small but non nil heritability might be regarded as corresponding to the mutual information carried by *UV* on P, with the frequencies for each value of E observed in this population being higher than the mutual information carried by *Genotype* on P.[28] The only difference between the diagrams of Fig. 2 and Pop_C is that *Genotype* and *UV* refer to potential difference makers in the former case and actual difference makers in the latter case.

[28] Had the variables *Genotype* and *UV* and *Cancer* been quantitative variables, from the measure obtained with Eq. (2), we would have seen that a small but non nil heritability of *Cancer* is equivalent to a situation in which the amount of variance of P explained by carrying out an intervention on *UV* is higher than when carrying out an intervention on *Genotype*. As seen above, it only corresponds partly to the 'range of influence' notion of causal specificity. This is because with quantitative variables, the values of two variables can be related in different ways—such as weakly or strongly—independently of range of influence. More on this point below.

If we now move to cases in which heritability is high (> 0.5), following a similar reasoning as with cases of low heritability, such cases correspond to situations in which phenotypic variation attributable to genotypic variation is superior to that attributable to environmental variation. A modification of the skin cancer example could be devised to show that a high heritability corresponds to a higher actual range of causal influence of the genotype on the phenotype than that of the environment.

It should be noted at this point that the fact that heritability is a ratio means that it cannot inform us about the absolute values of environmental and genotypic variance. Thus, a high heritability does not imply that a high variance in phenotype is mostly explained by a high variance in genotype. Rather it only implies that the variance in the phenotype associated with the variance in the genotype is larger than that associated with variance in the environment. This means that extreme cases in which there is no environmental variance in the population and *any* level of genetic variance different from 0 (even very small) will be cases in which heritability is 1. Thus, in terms of range of causal influence, cases of maximal heritability *only* imply a nil actual range of causal influence of the environment on the phenotype and a positive actual range of causal influence of the genotype on the phenotype.

Taking into account all the points of convergence between heritability and range of causal influence highlighted in this section, it is now time to attempt expressing the former in terms of the latter. Heritability corresponds, at least partly, to a measure of the actual range of causal influence of the genotype on the phenotype normalized by the range of actual variation of the phenotype, the latter of which is measured by the entropy of the phenotype. We thus have for nominal variables:

$$H^2 \equiv \frac{I(P; \widehat{G})}{H(P)} \tag{6}$$

where $I(P; \widehat{G})$ is the actual mutual causal information from G to P, and $H(P)$ is the entropy of P.[29] A similar definition for quantitative variables reads:

$$H^2 \equiv \frac{V(P : \widehat{G})}{V(P)} \tag{7}$$

where $V(P : \widehat{G})$ is the variance in P causally explained by G.[30]

To take an example with Pop_C, a heritability measure using Eq. (6) for nominal variables is $\frac{0.81}{2} = 0.405$ since $I(Cancer; \widehat{Genotype}) = 0.81\, bits$ and $H(Cancer) = 2\, bits$, assuming $G = Genotype$. and $P = Cancer$.

Before going further, something should be said about the relationship between Eqs. (6) and (7). In fact, the measure proposed in Eq. (6), because it applies to nominal variables, or to quantitative variables when the quantitative information of different

[29] Note that noise or error not attributable to the the variable G or the variable UV, assuming they are independent variables, corresponds, in information theoretic terms, to $H(P \mid \widehat{G}, \widehat{E})$, which measures the remaining entropy after having learned the value of the genotype and the environment (as it is measured, here by the variable UV), which is analogous to the *sum* of error terms in a regression model.

[30] I thank Arnaud Pocheville for proposing this measure.

values for a given variable are ignored, only captures one dimension of what a traditional measure of heritability or what Eq. (7) captures. In fact, if some particular interventions on genotype produce large changes in phenotype when compared to interventions of the same magnitude on the environment, or if one relationship is linear while the other is, say, quadratic, this will not be captured by the measure of Eq. (6). To be clear, as mentioned in Sect. 2, the differences made in such cases correspond to the dimension of causal strength which is not applicable to nominal variables since these are variables for which no comparisons of values can be made other than saying that these values are different.[31] Causal strength is captured by Eq. (7), however it is mixed with range of influence. These differences aside, both measures are similar, and one can recover the measure obtained in Eq. (6) from Eq. (7) provided some assumptions about the distributions of the variables are made (Garner and McGill 1956, pp. 226–227). For any practical purpose, as argued by Garner and McGill (1956), it is better, when dealing with quantitative variables, to use both the variance and information theory framework and compute the two types of measures. I believe this conclusion could be extended quite straightforwardly with regards to heritability.

With these relationships now established, I turn in the next section to the problem of locality and a few other objections to interpreting heritability causally.

5 Causation, locality and heritability

One of the main problems with interpreting heritability causally is what has been called the problem of locality (Lewontin 1974; Oftedal 2005; Taylor 2006, 2010; Sesardic 2005, p. 60): any measure of heritability obtained from one population cannot be compared to the heritability measure for the same phenotype in another population. In this section, I draw some links between the problem of locality and the concepts of invariance and stability proposed in the interventionist literature (see Woodward 2000, 2003, 2010; Pocheville et al. 2017; Griffiths et al. 2015). The problem of locality has been taken by some to be an important problem for interpreting heritability in causal terms. I argue for a more nuanced position which recognizes the problem of locality as an important one, but show that this problem concerns any causal relationship established in the special sciences. Finally, I discuss and defuse a few other problems that could arise from a causal interpretation of heritability.

We saw in the previous section, when presenting the potential/actual difference makers distinction, that two sorts of reasons can explain why any given heritability measure might be different in different populations, even if dealing with independent variables E and G. In fact, the frequencies for G and/or E might be different in different populations. I generalized this in terms of actual difference makers and argued that any amount of actual causal influence of one variable C on another variable O in one population might be different to the amount of actual causal influence of C on O in another population (assuming a constant background).

[31] As mentioned above, it does not make sense for the variable 'color', to ask whether the values 'red', 'yellow' and 'green' are superior or inferior to the value 'blue', if there is no way to reduce this variable to another quantifiable variable (such as the amount of a pigment, for instance).

Furthermore, besides what we considered so far as G and E, that is *genotype* and *UV* respectively, other factors in the background, and thus not considered, might have an effect on P (or on the relation between G and P or E and P). For instance, taking our cancer example, we could add in the background that the diet of individuals is different in different populations. A diet poor in antioxidant would reduce tolerance to solar UV exposure and increase the risk of skin cancer, whereas a diet rich in antioxidant would increase the tolerance. Given the difference in diet between different populations, it would be plausible to assume diet can have an influence on the probability to develop a skin cancer. Another example would be the differential use of sunscreen in different populations. These are cases of gene-environment statistical interaction[32] since background variables are variables in the environment. I will come back to the problem of gene-environment interactions and how it might be measured using information theory at the end of this section.

These two types of differences between populations embody (in part) the problem of locality identified by Lewontin (1974).[33] A consequence of locality is also that the same estimate of heritability obtained in two different populations, or at two different time points, might be the result of two very different underlying causal structures or distributions over the cause variable.

The problem of locality is one important reason which led Lewontin (1974) to conclude that heritability cannot generally be interpreted causally. Lewontin (1974), and with him a large number of authors (for a review see Sesardic 2005, Chap. 2), argues that an analysis of variance, upon which rest most heritability studies in humans, is not equivalent, in general, to an analysis of cause unless all causal relata involved are additive or nearly so. Others, while recognizing the problem of locality as important, have been less reluctant to regard it as a fatal problem (see Sesardic 2005; Lynch and Bourrat 2017).

As I show below, these two horns of the locality problem can be cashed out in terms of another dimension of causal relationships within the interventionist account, namely what has been called by Woodward (2003, Chap. 6) "invariance." Following Woodward, the invariance of causal relationship $C \rightarrow O$ is defined as the extent to which this relationship "remains stable or unchanged as various other changes occur" (2003, p. 239). Woodward (2003, 2010) distinguishes two kinds of invariance. The first one concerns the extent to which one can change the value of C without changing the causal relationship (or function) $C \rightarrow O$. Suppose that we established that C causes O and that the function mapping one value of C to one value of O is $O = f(C)$, for

[32] What appears as a statistical interaction, might causally be additive. We might consider a case in which antioxidant and sun exposure act additively on the probability to develop a skin cancer. Yet, without controlling for the background, this might appear as a statistical interaction.

[33] Lewontin writes for instance "That is, the linear model [upon which rests ANOVA] is a local analysis. It gives a result that depends upon the actual distribution of *genotypes and environments* in the particular population sampled. Therefore, the result of the analysis has a historical (i.e., spatiotemporal) limitation and is not in general a statement about *functional* relations. So, the genetic variance for a character in a population may be very small because the functional relationship between gene action and the character is weak for any conceivable genotype or it may be small simply because the population is homozygous for those loci that are of strong functional significance for the trait" (1974, p. 403, my emphasis for "genotypes and environments").

at least some intervention(s) on C.[34] The more invariant the relationship, the larger the number of interventions on C that will lead to changes of O that follow this relationship, and the more explanatory this relationship should be considered to be. To give an example, if developing a skin cancer as a result of solar UV exposure increases following a given function over a large range of values so that this relationship is considered highly invariant, then it seems reasonable to consider that the relationship is more explanatory than if this relationship only holds under a smaller range of solar UV exposures.

The second sort of invariance concerns the range of values that background variables of the relationship can take without changing the value of O or the properties of the relationship $C \rightarrow O$. Causal stability, or more precisely the lack of stability between G and P, corresponds to the notion of gene-environment interaction. The more unstable the relationship the higher the gene-environment interaction. Suppose that the function mapping one value of C to one value of O in a given background (materialized by the variable B) is $O = f(C)$, for at least some intervention(s) on C.[35] Under such conditions, the more stable this relationship, the larger the range of interventions on B that will keep the relationship $O = f(C)$, that is the mapping, unchanged. Using our example, if developing a skin cancer as a result of sun exposure, following a given pattern, is stable under a larger number of background conditions (for instance different diets) than say as a result of having a particular genotype, it seems reasonable to consider sun exposure as a more reliable cause of skin cancer than having a particular genotype because $UV \rightarrow Cancer$ is a more stable relationship than $Genotype \rightarrow Cancer$.[36] Pocheville et al. (2017), for clarification, call the first notion "invariance" and the second notion "stability." I follow suit.

In terms of range of causal influence, in cases where different causes (C and B) for an outcome (O) are independent, invariance and stability thus amount to an invariance in the range of causal influence across the possible range of values of a given C (which correspond to either G or E in heritability, assuming G and E are independent) and of B respectively. Pocheville et al. (2017) provide a suitable measure for stability, an information theoretic measure of invariance has yet to be developed.

This means that in heritability studies, the problem of locality boils down to a problem of low invariance and/or low stability for the relationships $G \rightarrow P$ and $E \rightarrow P$, since any change in one of these relationships will result in a change in the heritability estimate.

As a side note, when C and B interact, another measure of stability corresponds to the extent to which changing the causal range of influence between C and O depends on the value of B, assuming here that the mapping between C and O remains unchanged as B varies. Pocheville et al. (2017) provide a suitable measure based on another infor-

[34] Talking about functions for nominal variables can be problematic since with such variables one cannot map one number to another number. Instead of a mathematical function, think of a unified mechanism giving rise to the effect when the effect is invariant, as opposed to a large number of contributing mechanisms, when it has low invariance.

[35] I use the notion of function in a vernacular rather than mathematical sense here.

[36] Note that the stability of a relationship here does not inform us of the magnitude of the cause on the effect. Indeed, a causal relationship might be very stable, yet the magnitude of this relationship might also be very weak.

mation theoretic measure known as 'interaction information' different from the one just discussed to measure this sort of stability. This measure, applied in the context of heritability, can be interpreted as a measure of gene-environment interaction which—for the purpose of this article—I have assumed is nil. However, provided that there are well-known cases of gene-environment interaction in the literature (see Moffitt et al. 2005; Caspi et al. 2002), such a measure would be useful in this context. This example shows that the information-theoretic implementation of the interventionist account to investigate the links between causality and heritability is rich.

What should one conclude from having identified that the problem of locality for heritability corresponds to well identified concepts in the philosophy of causation literature? Woodward proposes that a low stability and/or a low invariance for a causal relationship should *not* lead to the conclusion that such a relationship is not causal, but there is a sense in which more invariant and more stable relationships have stronger explanatory power. Invariant/stable relationships provide better causal explanations. In fact, remember that relationships with low invariance and/or low stability all satisfy the minimal criterion of causation. They are thus all causal in this sense. I believe that applying this reasoning to heritability studies is warranted. Heritability estimates, even though they might vary between different studies are still indicators of a causal relationship between G and P in a population, assuming they rely on solid experimental data, or at least observational data with solid hypotheses enabling to derive a causal model from them (such as the representativity of the sample population). Furthermore, the extent to which the causal relationship between G and P is invariant/stable under intervention is an empirical question that will depend on the phenotype studied.

Another important reason why I regard Lewontin's conclusion, that most heritability studies are hopeless at capturing causation, as being too extreme is that if it was applied with parity to the relevant special sciences, that is sciences other than fundamental physics, one would have to conclude that many tests used in these disciplines do not and cannot in principle capture causation. In fact, a version of the problem of locality will be encountered for any system in which many variables (some of which will not be controlled) are causally involved in producing an effect. For instance, a recent study has shown that the replicability of experiments in psychological science might be below 40%[37], which led to what is known as a "replication crisis" (Open Science Collaboration 2015). One putative important factor explaining this low rate, among many other including methodological problems and publications bias, is a low stability of the relationships studied. Although replication studies attempt to have the same background variables when compared to an original study, there will inevitably be differences in the backgrounds (due to differences in uncontrolled variables) that might interact with the putative relationship being investigated. It is to be expected that the less stable the relationship, the less replicable its results will be.[38] This means that the low replicability in psychological science and elsewhere might not solely be due to the fact that a relationship investigated does not exist, but also that it is highly sen-

[37] Replicability means here that if a study had a statistically significant result, its replication also had a statistically significant result.

[38] Another independent problem is whether the experimental conditions, or the conditions under which studies are conducted, represent natural conditions, which can also be cashed out in terms of differences in backgrounds.

sitive to the background in which it was originally established. Applying Lewontin's conclusion about heritability estimates to these studies would lead to the conclusion that interpreting many results from psychological science causally is misguided.

I do not think that it is what one should conclude. Rather, a more reasonable approach, in my view, is that any causal claim made from one single study involving a complex system should be taken with a grain of salt, and one should attempt to find evidence that could corroborate or contradict this causal claim as well as the stability of the relationship. This clearly is the main function of replication, which is one important, yet often neglected, motor of scientific progress. Thus, facing a given heritability estimate (whether it originates from observed or experimental data) one ought to question the robustness of this finding in order to evaluate to what extent both $G \rightarrow P$ and $E \rightarrow P$ vary for different values of G and E, different frequencies for these values, and in different backgrounds, and to what extent the results obtained can be extrapolated beyond the range of background conditions in which it has been established. At the very least, any causal interpretation of heritability ought to be contextualized by giving as much information as possible about the background (the population) against which the causal claim is made. But heritability studies are by no means the sole case in which such precautions should be taken. This is a general methodological point that applies to all scientific disciplines. In short, the fault with causal interpretations of heritability estimates lies mostly in what people have made of them rather than in the estimates themselves. There is indeed evidence that consideration of agency, moral responsibility, and the type of phenotype investigated, influence reasoning about whether variation in this phenotype is the outcome of variation in the environment or in the genes (see for instance Alicke 1992; Lynch 2017).

Before concluding, it should be noted that interpreting heritability in the way I did here may also shed some light on a possible criticism that a causal interpretation of heritability might encounter, which is in the context of behavioral genetics, namely that because everything is heritable (Turkheimer's 1998; 2000 first law of behavioral genetics), interpreting heritability causally explains nothing. It is in fact true that most traits, including traits such as religiosity, marital status, or political opinions, which would typically be considered as pertaining to the cultural domain rather than the biological one, are all heritable to some extent (Polderman et al. 2015). The worry here is that, if any trait turns out to be heritable with approximately the same value, then its discriminatory or explanatory power will be diminished, in the same way that invoking God to explain any event happening in the life of someone is not explanatory.

One way to respond to this criticism is that my causal interpretation leads to the view that heritability is a *relative* measure of genotypic causal influence on phenotype in a *particular context*. That it stresses that the terms 'relative' and 'particular context' provide an antidote against naive causal interpretations. Yes, all these traits are causally influenced by the genotype, but that does not permit to answer the following questions: Is the relationship the same in all contexts? What if there is more variation in the environment? Does an intervention in the environment or in the individuals of a population influence the phenotypic outcome? In sum, a naive causal interpretation of heritability cannot appreciate that, in spite of a high heritability, the background or context (including the environment) can be a much more important source of variation

(which furthermore might be hidden given a lack of actual variation) in phenotype than is the genotype.

In many regards, my causal interpretation vindicates Lewontin's recommendation to switch from thinking about the relations between genotype, the environment and phenotype in terms of heritability to thinking about these relations in terms of norms of reaction. A positive heritability estimate (assuming confounds have been eliminated) establishes that there is a causal link between genotype and phenotype, but it says nothing about the magnitude of this link, nor whether it breaks down easily. It does not follow however, that because heritability has some important limitations when it comes to a causal interpretation, it is not a valid tool for causal interpretations, especially when the limitations are made explicit and no genuine alternative exists.

6 Conclusion

In this paper I have shown that heritability, when interpreted causally within the interventionist account of causation, is partly commensurate with the notion of causal specificity *qua* range of causal influence proposed by Woodward (2010).[39] More particularly, starting from the formalization in information-theoretic terms of this dimension of causation proposed by Griffiths et al. (2015) and the notion of an actual difference maker proposed by Waters (2007), I argued that heritability amounts, at least partly, to an *actual* range of causal genotypic influence normalized against an *actual* range of phenotypic variation within a population. I then showed that any interpretation of a given heritability estimate is prone to a form of the problem of locality identified by Lewontin (1974). I linked this problem to the notions of invariance and stability encountered in the interventionist literature. I argued that the problem of locality with heritability, insofar as it amounts to a low stability and/or low invariance of the causal relationships between, on the one hand, genotype and phenotype, and on the other hand, environment and phenotype, is only one exemplar of what can be regarded as a ubiquitous problem in science. A natural next step in the overall project of linking the interventionist account to heritability would be to provide an analysis of different heritability *estimates* using the interventionist account to establish how different types of estimates fare with respect to Woodward's stability and invariance and ultimately how causally explanatory they are.

In the context within which heritability estimates have been used to justify and reify differences (such as differences in IQ) observed between groups of humans with different ethnic backgrounds, I believe, following Lynch and Bourrat (2017), that a precise characterization of the scope and limits of interpreting heritability causally represents a way forward towards eliminating harmful political biases.

Acknowledgements I am thankful to the Theory and Method in Biosciences group at the University of Sydney for feedback on a previous version of the manuscript. I thank in particular Arnaud Pocheville for discussion and insightful comments, and Stefan Gawronski who proofread the final manuscript. This research was supported by a Macquarie University Research Fellowship and a Large Grant from the John Templeton Foundation (Grant ID 60811).

[39] It is only partly commensurate, for reasons detailed at the end of Sect. 4 with respect to the fact that the measure of range of causal influence, proposed by Griffiths et al, does not consider quantitative variables.

References

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378. https://doi.org/10.1037/0022-3514.63.3.368.

Begg, C. B., Orlow, I., Hummer, A. J., Armstrong, B. K., Kricker, A., Marrett, L. D., et al. (2005). Lifetime risk of melanoma in CDKN2A mutation carriers in a population-based sample. *JNCI: Journal of the National Cancer Institute*, *97*(20), 1507–1515. https://doi.org/10.1093/jnci/dji312.

Bell, A. E. (1977). Heritability in retrospect. *Journal of Heredity*, *68*(5), 297–300. https://doi.org/10.1093/oxfordjournals.jhered.a108840.

Bourrat, P. (2015). How to read 'heritability' in the recipe approach to natural selection. *The British Journal for the Philosophy of Science*, *66*(4), 883–903. https://doi.org/10.1093/bjps/axu015.

Bourrat, P. (2019). Evolution is About Populations, But Its Causes are About Individuals. *Biological Theory*, *14*(4), 254–266.

Bourrat, P. (2019a). On Calcott's permissive and instructive cause distinction. *Biology & Philosophy*, *34*(1), 1. https://doi.org/10.1007/s10539-018-9654-y. 00001.

Bourrat, P. (2019b). Variation of information as a measure of one-to-one causal specificity. *European Journal for Philosophy of Science*, *9*(1), 11. https://doi.org/10.1007/s13194-018-0224-6. 00001.

Bourrat, P. (accepted) Causation and SNP Heritability'. Philosophy of Science.

Bourrat, P., & Lu, Q. (2017). Dissolving the missing heritability problem. *Philosophy of Science*, *84*(5), 1055–1067.

Caspi, A., McClay, J., Moffitt, T. E., Mill, J., Martin, J., Craig, I. W., et al. (2002). Role of genotype in the cycle of violence in maltreated children. *Science*, *297*(5582), 851–854. https://doi.org/10.1126/science.1072290.

Downes. S. M. (2009). *Heritability*. Stanford Encyclopeadia of Philosophy.

Falconer, D. S., & Mackay, T. F. (1996). *Introduction to quantitative genetics* (4th ed.). Essex: Longman.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(02), 399–433.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Picconi, O., Boyle, P., et al. (2005a). Meta-analysis of risk factors for cutaneous melanoma: II. Sun exposure. *European Journal of Cancer*, *41*(1), 45–60. https://doi.org/10.1016/j.ejca.2004.10.016.

Gandini, S., Sera, F., Cattaruzza, M. S., Pasquini, P., Zanetti, R., Masini, C., et al. (2005b). Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors. *European Journal of Cancer*, *41*(14), 2040–2059. https://doi.org/10.1016/j.ejca.2005.03.034.

Garner, W. R., & McGill, W. J. (1956). The relation between information and variance analyses. *Psychometrika*, *21*(3), 219–228. https://doi.org/10.1007/BF02289132.

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection*. New York: Oxford University Press.

Griffiths, P. E., & Neumann-Held, E. M. (1999). The many faces of the gene. *Bioscience*, *49*, 656–662.

Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., & Knight, R. (2015). Measuring causal specificity. *Philosophy of Science*, *82*(4), 529–555. https://doi.org/10.1086/682914.

Griffiths, P. E., & Stotz, K. (2013). *Genetics and philosophy: An introduction*. New York: Cambridge University Press.

Hayward, N. K. (2003). Genetics of melanoma predisposition. *Oncogene*, *22*(20), 3053–3062. https://doi.org/10.1038/sj.onc.1206445.

Jacquard, A. (1983). Heritability: One word, three concepts. *Biometrics*, *39*, 465–477.

Korb, K. B., Nyberg, E. P., & Hope, L. (2011). A new causal power theory. In F. Russo, J. Williamson, & P. M. Illari (Eds.), *Causality in the sciences* (pp. 628–652). Oxford: Oxford University Press.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197. https://doi.org/10.2307/2678389.

Lewontin, R. C. (1974). Annotation: The analysis of variance and the analysis of causes. *American Journal of Human Genetics*, *26*(3), 400–411.

Lu, Q., & Bourrat, P. (2018). The evolutionary gene and the extended evolutionary synthesis. *The British Journal for the Philosophy of Science*, *69*(3), 775–800. https://doi.org/10.1093/bjps/axw035.

Lynch, K. E. (2017). Heritability and causal reasoning. *Biology & Philosophy*, *32*(1), 25–49. https://doi.org/10.1007/s10539-016-9535-1.

Lynch, K. E., & Bourrat, P. (2017). Interpreting heritability causally. *Philosophy of Science*, *84*(1), 14–34. https://doi.org/10.1086/688933.

Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits* (Vol. 1). Sunderland, MA: Sinauer.
Moffitt, T. E., Caspi, A., & Rutter, M. (2005). Strategy for investigating interactions between measured genes and measured environments. *Archives of General Psychiatry*, *62*(5), 473–481.
Northcott, R. (2006). Causal efficacy and the analysis of variance. *Biology and Philosophy*, *21*(2), 253–276. https://doi.org/10.1007/s10539-005-8304-3.
Northcott, R. (2008). Can ANOVA measure causal strength? *The Quarterly Review of Biology*, *83*(1), 47–55. https://doi.org/10.1086/529562.
Oftedal, G. (2005). Heritability and genetic causation. *Philosophy of Science*, *72*(5), 699–709. https://doi.org/10.1086/508126.
OSC (Open Science Collaboration). (2015). Estimating the reproducibility of psychological science. *Science349*(6251):aac4716. https://doi.org/10.1126/science.aac4716.
Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.
Pocheville, A., Griffiths, P. E., & Stotz, K. (2017). Comparing causes—an information-theoretic approach to specificity, proportionality and stability. In H. Leitgeb., I. Niiniluoto., E. Sober., & P. Seppälä (Eds.), *Proceedings of the 15th congress of logic, methodology and philosophy of science*. London: College Publications.
Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., et al. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, *47*(7), 702–709. https://doi.org/10.1038/ng.3285.
Sarkar, S. (1998). *Genetics and reductionism*. Cambridge: Cambridge University Press.
Sesardic, N. (2005). *Making sense of heritability*. Cambridge: Cambridge University Press.
Sterelny, K., & Griffiths, P. E. (1999). *Sex and death: An introduction to philosophy of biology*. Chicago: University of Chicago Press.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677.
Stewart, B. W. K. P., & Wild, C. P. (2014). *World cancer report 2014*. Geneva: WHO.
Stone, J. V. (2015). *Information theory: A tutorial introduction*. Sheffield: Sebtel Press.
Tabery, J. (2014). *Beyond versus: The struggle to define the interaction of nature and nurture*. Cambridge, MA: MIT Press.
Tal, O. (2009). From heritability to probability. *Biology & Philosophy*, *24*(1), 81–105. https://doi.org/10.1007/s10539-008-9129-7.
Tal, O. (2012). The impact of gene-environment interaction and correlation on the interpretation of heritability. *Acta Biotheoretica*, *60*(3), 225–237. https://doi.org/10.1007/s10441-011-9139-8.
Taylor, P. (2006). Commentary: The analysis of variance is an analysis of causes (of a very circumscribed kind). *International Journal of Epidemiology*, *35*, 527–531.
Taylor, P. (2009). Perspectives from plant breeding on Tal's argument about the weight of genetic versus environmental causes for individuals. *Biology & Philosophy*, *24*(5), 735–738. https://doi.org/10.1007/s10539-009-9162-1.
Taylor, P. (2010). Three puzzles and eight gaps: What heritability studies and critical commentaries have not paid enough attention to. *Biology & Philosophy*, *25*(1), 1–31. https://doi.org/10.1007/s10539-009-9174-x.
Taylor, P. J. (2012). A gene-free formulation of classical quantitative genetics used to examine results and interpretations under three standard assumptions. *Acta Biotheoretica*, *60*(4), 357–378. https://doi.org/10.1007/s10441-012-9164-2.
Turkheimer, E. (1998). Heritability and biological explanation. *Psychological Review*, *105*(4), 782.
Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, *9*(5), 160–164. https://doi.org/10.1111/1467-8721.00084.
Visscher, P. M., & Goddard, M. E. (2019). From R.A. Fisher's 1918 paper to GWAS a century later. *Genetics*, *211*(4), 1125–1130. https://doi.org/10.1534/genetics.118.301594.
Visscher, P. M., Hill, W. G., & Wray, N. R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, *9*(4), 255–266.
Waters, C. K. (2007). Causes that make a difference. *The Journal of Philosophy*, *104*, 551–579.
Weber, M. (2006). The central dogma as a thesis of causal specificity. *History and Philosophy of the Life Sciences*, *28*, 595–609.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science*, *51*(2), 197–254.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.
Woodward, J. (2013). Causation and manipulability. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Moscow: Winter.
Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1–24.
World Health Organization. (2002). Global solar UV index: A practical guide. World Health Organization, Geneva, Switzerland, oCLC: 51304373.
Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs. *Proceedings of the National Academy of Sciences*, *6*(6), 320–332. https://doi.org/10.1073/pnas.6.6.320.
Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245–280. https://doi.org/10.2307/2185535.
Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, *42*(7), 565–569.