Policy recommendations from causal inference in physics education research

M. B. Weissman

Department of Physics, University of Illinois at Urbana-Champaign,
1110 West Green Street, Urbana, Illinois 61801-3080, USA

Sound educational policy recommendations require valid estimates of causal effects, but observational studies in physics education research sometimes have loosely specified causal hypotheses. The connections between the observational data and the explicit or implicit causal conclusions are sometimes misstated. The link between the causal conclusions reached and the policy recommendations made is also sometimes loose. Causal graphs are used to illustrate these issues in several papers from *Physical Review Physics Education Research*. For example, the core causal conclusion of one paper rests entirely on the choice of a causal direction although an unstated plausible alternative gives an exactly equal fit to the data.

I. INTRODUCTION

A central goal of most physics education research (PER) is to find ways to teach better, i.e., to improve educational outcomes. That requires determining the probable results of several possible courses of action, so that value judgments can be applied to choose the best policies. Problems of the form "If we do this vs that, what will be the probable difference in results?" are by definition problems in counterfactual or potential outcome causation. (See Ref. [1] for a standard text, or Ref. [2] for an up-to date technical review.) Techniques for drawing causal conclusions from randomized controlled trials are generally understood, but the traditional methods for drawing causal inferences from observational data (e.g., multiple regression) are often inappropriate or misapplied [1,2]. As PER has moved to tackle increasingly challenging problems, the causal inference challenges have grown. "Correlation does not imply causation" is a truism that is easier to state than to follow in practice.

The purpose of this paper is to illustrate the need for better causal analysis in PER by using a few instances of papers published in this journal in which questionable procedures were used either to draw causal inferences from data or to draw policy recommendations from causal inferences or both. It is motivated by exchanges both with statisticians discontented with the level of causal reasoning in many social sciences and with colleagues in PER who were both uneasy about their lack of familiarity with modern techniques and curious about how well some recent PER work would hold up under scrutiny. I've chosen several papers by leaders in PER to illustrate what are said to be some common issues.

I call attention to these issues because improvement is possible. In recent decades more valid methods of causal inference have been developed [1,2] with applications in epidemiology [2], biology [3], public health [4], economics [5], psychology [6], sociology [7], political science [8], education [9], and other fields. Although here I will focus on instances in which causal conclusions rely on unstated questionable assumptions, the new methods allow causal conclusions to be made from complicated data so long as reasonable explicit assumptions are made [1,2].

I will use only the elementary level of analysis with which I am familiar. Other than encouraging more study of causal inference, I will make no educational recommendations, but will suggest an editorial change for the field.

Before presenting an admittedly blunt critique of the causal reasoning in these PER papers, I should acknowledge that much of the correlational data and analysis presented in them may well prove useful. My point is that invalid causal conclusions can be found even in papers treating correlational results carefully. Some additional data, ideally obtained by measuring effects of various interventions, will be needed in order to draw reliable causal conclusions to be used in making plans.

II. BACKGROUND ON CAUSAL METHODS

Graphical representations of causal patterns [1], developed especially by Pearl and co-workers, are perhaps the most accessible of the newer methods and will suffice for this paper. Graphs are used to represent the most important feature of causal relations, that changing some quantities leads to changes in only some others. Graphs called

directed acyclic graphs (DAGs) represent this one-way feature by using arrows to show the causal directions [1]. A change in any variable only affects variables downstream from it in the DAG, and there are no causal loops. For physicists, the most familiar such relation is that changes propagate only forward in time, at least on the classical macroscale. A DAG itself does not give the functional form, magnitude, or even sign of the causal effects, but it does encode conditional independence relations, i.e., which interventions can have no effect on which other variables. DAGs are particularly useful for visualizing which variables should be controlled for, which should not be controlled for, and which create less tractable problems when estimating causal relations from correlations [1].

A bidirectional arrow is sometimes used as a shorthand representation of an unmeasured variable with no arrows going in and arrows going out to two measured variables. Graphs using that shorthand are called acyclic directed mixed graphs (ADMGs). Their bidirectional arrows should not be mistaken for bidirectional causation.

The papers to be discussed here assume that the relations among the variables can be well approximated as linear. Then a linear structural equation model (SEM) can explain the correlations among the variables in terms of the causal structure expressed in the ADMG [10]. Each causal arrow gets a coefficient showing the *direct* effect of changes of the tail variable on the head variable. The *net* causal effect of one variable on another is obtained by summing the products of the coefficients along each path leading from the first variable to the second [11].

The directions of the arrows provide essential information needed to determine policy choices, since making a change in some quantity only affects variables downstream [1]. You can make a dog wag its tail by giving it a treat, but you cannot put a treat in its mouth by wagging its tail. Although words like "effect" and "impact" are often used to describe correlations while suggesting causation, the statements "Giving treats has a tail-wagging effect" and "tail-wagging has a treat-giving effect" do not mean the same thing. The difference would appear in the direction of the arrow between "treat" and "wag." The choice of graphs has major practical importance because they capture this causal distinction.

If the relevant ADMG is not already known, SEMs can only play a limited role in inferring causation from correlations [11]. Correlational data can rule out some ADMGs, e.g., ones with no pathways connecting variables that are significantly correlated, but that is far short of determining which ADMG should be used as the framework for the SEM. Often SEMs based on many different graphs can approximately fit the same dataset. Furthermore, there are theorems showing that the ADMGs can be divided into Markov equivalence classes, each of whose members fit any given data set *exactly* equally well, with the same number of adjustable parameters [1–3,12]. There are known rules for generating equivalence classes of graphs [1,12]. On their own, SEMs provide evidence for choosing among equivalence classes but not for choosing graphs within a class. The choice among such equivalent graphs should be left open unless there is other evidence such as time order, other types of data, or firm knowledge about mechanisms, although intuition may guide tentative beliefs [11].

In the first two papers described, particular graphs are chosen as the basis for SEMs but without sufficient reason to exclude other graphs with very different causal implications. I will use alternative members of equivalence classes to illustrate rigorously how big those causal differences can be, but not to advocate any particular causal model. A third paper uses SEMs carefully, but then draws strong causal conclusions about unmeasured variables outside the range described by its graph.

III. CORRELATIONS DO NOT DETERMINE CAUSAL GRAPHS

I shall discuss two papers [13,14] that say little about how the directions of graphical arrows were chosen, and whose causal conclusions turn out not to be justified by their data. These papers use SEMs as if they were a nearly stand-alone method for discovering causal patterns. [13,14].

These papers [13,14] look at traits like "interest" rather than events. Gradually developed traits can be unrolled into functions of time. It is then inappropriate to represent a snapshot in time of a set of such traits by *any* ADMG since effects generally have been flowing *both ways* between past values of the traits [2]. (Fortunately, methods have been developed for inferring causal relations in complicated time-dependent longitudinal data [2,15].) Thus when I suggest alternatives to the causal interpretations used in these papers, I do not mean to imply that these alternatives are better and certainly not that they are right, since it is unlikely that any ADMG correctly represents the causal relations between snapshots of prolonged traits. Although this is a fundamental problem for the sorts of SEM models to be discussed, I will confine the discussion to narrower methodology issues.

A. Gender

One paper [13] relates gender to questionnaire-based measures of attitudes toward physics, "…to explore which motivational factors cause changes in other factors…" [13]. The attitudes are grouped into four clusters: "interest," "perceived recognition," "competency belief," and "physics person," with the latter three abbreviated here as recognition, competence, and identity for notation compatible with [14]. The linear SEM model used [13] is based on the DAG shown in Fig. 1. Since gender almost always precedes the attitudes examined, it can be assumed to have no
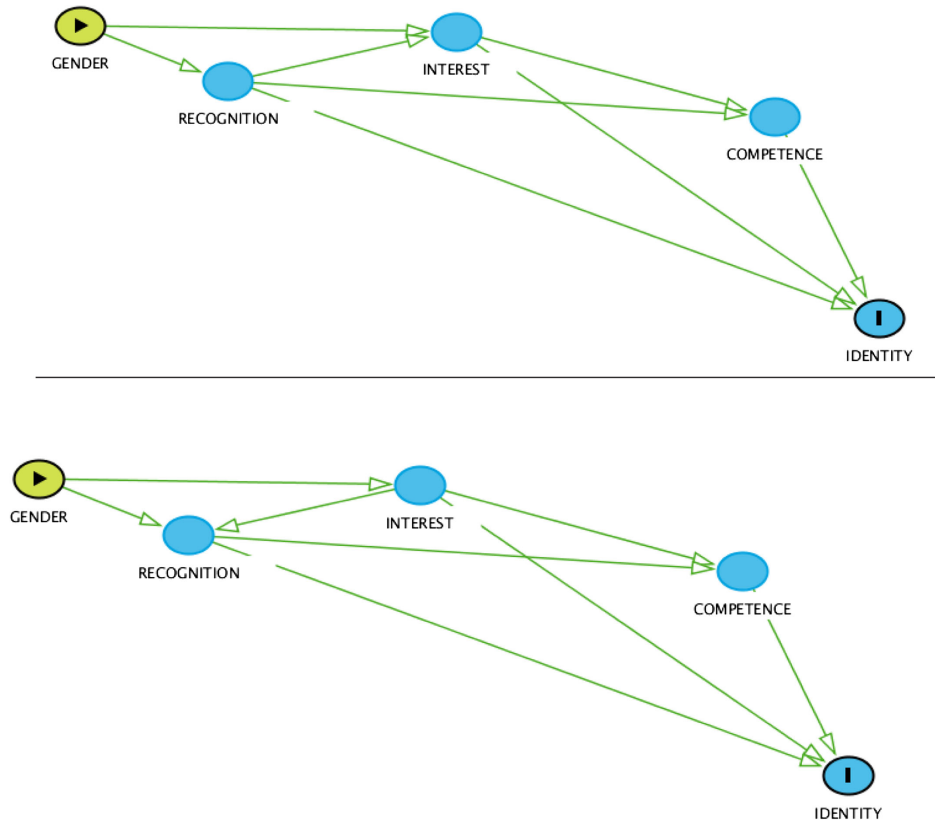
FIG. 1. The top graph is the DAG from Ref. [13], using abbreviated variable names. The bottom one is the Markov equivalent graph described in the text. The graphs were drawn using the online DAGitty tool.

incoming arrows. (Selection effects on who gets into the sample of calculus-based physics students could complicate that relation [16], but I shall not explore such effects here.)

This graph was chosen by "dropping connections or variables of low strength" from the saturated model in which all variables are connected, although the minimum strength for inclusion is not specified [13]. The two connections dropped were from gender to competence and to identity. All six connections between the four core attitudinal variables are retained.

How the arrow *directions* were chosen is described only by "…we only used the suggestions that were theoretically plausible." [13] Markov equivalence rules say the arrows between interest and recognition and between identity and competence can each independently be assigned either direction, leaving four equivalent DAGs [1,12]. (Either of these two arrows could also be replaced by a bidirectional arrow giving nine equivalent ADMGs [12].)

Theoretical plausibility seems inadequate for deciding the directions of the arrows, especially between interest and recognition. Do we really know that students become interested in areas for which they perceive recognition more than that they perceive recognition for areas in which they have been interested? The equivalent graph shown in Fig. 1 obtained if the arrow from recognition to interest

were reversed would have substantially changed meaning. The coefficient between interest and recognition would increase slightly, from 0.64 to 0.67. The coefficient from gender to interest would increase from 0.16 to 0.33, and the coefficient from gender to recognition would fall from 0.27 to 0.05. The qualitative impression from this equivalent graph would then be that gender differences in interests are most important, rather than gender differences in perceived recognition. The data do not provide guidance on which choice is better.

The very small coefficient for the Gender → Recognition arrow in this equivalent graph raises the question of whether it would meet the unspecified criteria for "dropping connections or variables of low strength", since it is only one-third as large as the smallest coefficient included. [13] Based on the sample size of 559, [13] the 95% confidence interval for this coefficient would be approximately (−0.05, 0.15), so that conventionally one would expect the coefficient to be dropped from the model as statistically insignificant. If it were dropped, then in the resulting new equivalence class all gender effects would be mediated through interest. This equivalence class also includes some DAGs for which recognition has no causal effects on other traits.

The claim of the title of Ref. [13], that females "do not identify with physics [because] they do not think others see

them that way" thus lacks support in the data. The data shown are *exactly* as compatible with a very small and statistically insignificant effect of gender on perceived recognition as they are with the moderate-size effect presented. Interpretations in which the gender → recognition and recognition → identity effects are both zero are disfavored only by that insignificant coefficient. One simply cannot tell from these data to what extent the paper's causal claims are correct.

The broad policy recommendations, centered on making the physics classroom gestalt more supportive, did not require careful identification of causal relations among the attitudinal variables [13]. An overlapping group of authors did report that a controlled intervention [17] encouraging effort when faced with difficulty and facilitating supportive interactions among students gave important improvements in physics course performance for females and perhaps also for nonwhite students, although the effect did not reach conventional statistical significance for the small nonwhite sample [17]. It did not include a survey of the traits studied in Refs. [13,14], so it is hard to tell which, if any, of these traits were most affected. This experiment serves as a reminder of how much more straightforward it is to find out what works from interventions than from static correlations.

B. Out-of-class activities

The effects of out-class science and engineering activities (OCSE) on student attitudes were explored in a recent paper for which "…the primary goal of the current analysis is determining the impact of OCSE activities on physics identity", i.e., estimating a causal effect [14]. The abstract raises a policy question based on causal effects: "Understanding the influence of students' science and engineering experiences on career choices is critical in order to improve future efforts…" [14]. The abstract also reaches a causal conclusion "out-of-class science and engineering activities have the largest influence on physics performance/competence beliefs…" [14]. Although at points noncausal associative wording is used, the body of the paper is peppered with causal conclusions, e.g., "Recognition beliefs, while having the largest impact on overall identity…" and "…physics identities have less impact on their career choice… "[14]. In one case a conclusion is drawn explicitly about the expectation of what would happen if something is done: "…if performance/competence beliefs are developed in isolation from recognition beliefs and interest, a student is not more likely to develop an overall physics identity" [14].

The causal conclusions are based on coefficients in an SEM relating OCSE to the same questionnaire-based measures of attitudes toward physics used in Ref. [13], here called interest, recognition, performance or competence, and physics identity. (I again use competence and identity as abbreviations for the latter two.) The ADMG
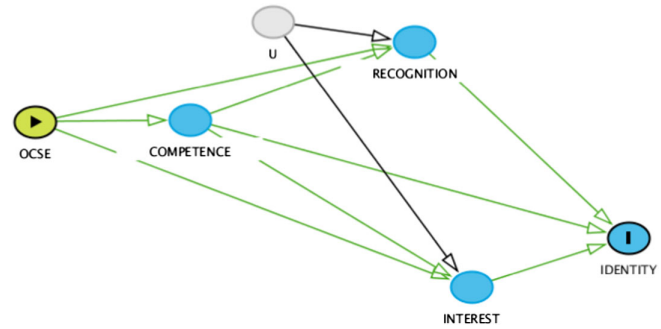

FIG. 2. This is the DAG form of the ADMG used in Ref. [14], with their bidirectional arrow shown instead as an explicit unmeasured U. The variable "Physics Career Choice," connected only by an arrow from identity, is omitted for simplicity.

connecting these variables [14] is redrawn in DAG form as Fig. 2 after converting the double arrow of the original to an explicit unmeasured variable U.

Reference [14] does not mention any criteria used in its choice of graphs, but says that the graph is based on prior work [18]. That work says "we need to caution that the word effect in the following exploration of the results cannot imply causality in a strict sense…" and "… it is important to keep in mind that this study is correlational, so causality is not certain." The prior work also does not explore the class of graphs that are equivalent to the one they chose [18], which include the same nine ADMGs connecting interest, recognition, and competence we saw for Ref. [13].

The graph chosen in Ref. [14] modifies the one in Ref. [18] by adding OCSE, with three arrows coming out and none going in. Those directions may not represent how the world works. If interested students often choose to do extra science activities, for example, interest may have more causal effect on OCSE than OCSE has on interest.

If the causal assumptions implicit in the graph chosen are correct, then the "primary goal," "determining the impact of OCSE activities on physics identity" [14], is simply given by the unconditional regression coefficient of identity on OCSE, since no confounders are present [1]. Neither this simple relation nor the resulting causal coefficient is given in the results [14].

The quality of the SEM fit to the data is discussed but not in comparison to fits using other graphs [14]. The graph choice could not have been based on some standard trade-off between fit quality and number of parameters, e.g., the Akaike information criterion [19], because elementary DAG rules [1] say that reversing the arrow from competence to OCSE gives a graph in the same equivalence class. That reversal, however, would make the unconditional regression of Identity on OCSE *not* equal to the impact of OCSE activities on physics identity and thus would change the implied estimate of "…the primary goal …the

impact of OCSE activities on physics identity". There may also be other equivalence classes that fit better [14].

The ADMG bidirectional arrow here has sufficient properties to allow replacement with either unidirectional arrow while preserving equivalence [12], giving many graphs in the same equivalence class. Even if we leave identity fixed as a child of all but OCSE, there are 24 DAGs obtained from the permutations of the other labels in the saturated graph of the other four variables. (In addition, there are many equivalent ways of substituting one or two bidirectional arrows for unidirectional ones.) Two of the equivalent DAGs would have reversed directions of all three arrows from OCSE. For those equivalent graphs "the impact of OCSE activities on physics identity" would be identically zero.

Despite the plethora of equivalent graphs available, there is one indication that the ADMG chosen does not represent causality very well. The authors remark upon the "surprising" negative sign of the regression coefficient for the direct effect of competence on identity [14]. That interpretation results from inappropriately treating the negative SEM coefficient found in a causally unmotivated graph as if it were a causal coefficient describing what would happen "…if performance/competence beliefs are developed in isolation from recognition beliefs and interest….". As an association, however, the negative coefficient is not surprising. A student who is confident in their abilities but still neither has interest in a field nor has sought recognition for it probably finds that field unattractive for some other reason and thus does not identify with it.

The surprising negative coefficient provides an example in which SEM results can help guide the formulation of causal hypotheses using graphical methods. This verbal description could be translated to a graph description by a slight modification of the graph used, which already includes an unmeasured cause U affecting both interest and recognition. U could be any number of causes that would be likely to have the same sign of effect on identity as on interest and recognition. If there should also be a causal arrow from U to identity with the same sign of coefficient as the arrows from U to interest and recognition, the regression coefficient obtained for identity on competence in the graph without the U → identity arrow would differ from the direct causal coefficient by an undetermined amount of negative collider stratification bias, since interest and recognition are colliders between U and competence [20].

What sort of traits could contribute to the implicit U? Many factors are believed to affect how attractive a scientific field is to some student. [21] Anecdotally, it is easy to think of many friends who were good or even brilliant physics students but who knew early on that their strongest interests lay elsewhere. One previous study found a negative correlation between desiring a career with interpersonal interactions and identifying as a physics person [22]. An experimental study from an overlapping group showed that discussing the underrepresentation of women in physics leads to more of them expressing interest in physics careers [23]. Closer integration of such experimental results with attitudinal survey data might produce clearer and more useful predictions. These informal suggestions are just meant as a reminder that a combination of explicit prior ideas about causality together with SEM results can suggest causal models with testable predictions [11].

The paper's conclusion draws policy recommendations based on what the effects would be of "Modifying programmed activities to better support recognition beliefs and interest…" or "…recognizing students in programmed OCSE activities …" [14]. The recommendation to "… [ensure] that activities are not only fun and engaging but also provide challenge. Students need to be provided with sufficient guiding support…" is not particularly reliant on the causal analysis [14].

C. Comparing (a) and (b)

Both studies [13,14] include all six connections among the four shared core attitudinal variables. One connection is represented with a unidirectional arrow in one study [13] and with a bidirectional arrow representing the influence of an implicit unmeasured variable in the other study [14]. Two connections are represented with arrows pointing opposite directions in the two studies. Three are represented with arrows in the same direction in both studies. One of those three arrows has opposite signs of coefficients in the two studies. Of the six connections only two are represented with arrows of the same type with the same sign of coefficient. *It would be hard to argue that the two graphs representing these variables are close to converging toward a shared causal picture.*

No amount of linear SEM analysis can decide which causal pattern is best without further inputs. One obvious type of input would be longitudinal data. It may be that no ADMG could represent the causal relations among the traits unless each trait were unrolled into a time series. Specific interventions in randomized controlled trials would provide stronger evidence. For example, Ref. [24] showed that an intervention designed to increase interest in secondary-school science also led to an increase in actual performance among the students expected to have the lowest performance. That is likely to in turn have led to an increase in the sense of competency measured in these surveys, consistent with the arrow direction chosen in Ref. [13] but opposite to that in Ref. [14]. For prolonged traits both directions can coexist.

IV. CAUSATION WITHIN AND OUTSIDE A GRAPH

Neither of the papers discussed above adequately consider the variety of different causal relations that might exist

among the measured traits. I now turn to a paper that uses SEM analysis properly for the traits measured, but then draws strong causal conclusions about unmeasured causes not explored in the SEM analysis [25]. I will illustrate the inability of the data to support the causal claims by showing a DAG including those unmeasured causes.

The paper [25] claims that the differences between major demographic groups in introductory college physics test scores are due to differences in precourse "preparation." The explicit data analysis given shows that when college exam scores of individuals are predicted from a combination of precourse scores on math ACT/SAT tests and the physics "concept inventory" (CI) exams, then the coefficients of demographic variables included in the predictive model are not large enough to consider statistically significant in this sample, with the possible exception of "first generation" status. The point estimates of these demographic coefficients are small for practical purposes. [25] That shows that the incoming test combination (unlike either incoming test on its own) is an approximately unbiased predictor of college physics exam scores with respect to the demographic variables considered. That result could be useful in evaluating interventions designed to reduce demographic disparities, helping in making comparisons of outcomes in non-randomized comparison groups.

On the way from this valid result toward policy suggestions, however, the paper makes a clear causal claim about what the effects of an action would be: "This work shows that creating instruction that enhances the success of every student across the full range of incoming preparations is also the solution to eliminating gaps in the performance across demographic groups" [25]. That conclusion sounds reasonable and may be supported by other work, but this paper itself has no data on different teaching methods, so it cannot show what their effects would be on different students.

The main result was interpreted to mean that "preparation gaps" are responsible for the demographic differences in college scores [25]. "Preparation" is not defined in the article, but other traits such as "test anxiety and stereotype threat" are mentioned as possible alternatives to it, indicating that preparation is not meant as a catch-all term for all precourse traits that might affect the test results [25]. If preparation is nonetheless meant as a catch-all for all traits that actually reflect facility with the material, then the description of the results in terms of it is somewhat tautological. That does not seem to be the meaning intended in the paper, which describes the precourse tests as "admittedly crude proxies of incoming preparation" [25]. Thus, the measured "preparation" accounts for only a fraction of the individual precourse score variance, with the rest of the score variance coming from other causes. This strong meaning, in which preparation is not the only possible cause but is observed to be the only important

one for demographic differences, was conveyed without ambiguity in a subsequent editorial entitled "It's Not 'Talent,' it's 'Privilege'" by the senior author, explicitly claiming that Ref. [25] shows that talent plays little role in the individual score differentials [26].

That stronger version of the interpretation is not justified by the data. Any fairly stable individual trait measurable by these exams would give the results observed, so long as its effects on both the precourse and college tests were fairly large and about the same in each demographic group, and other independent traits were approximately demographically balanced. Talent, discussed but rejected in the subsequent editorial [26], or interest, discussed in the papers described above [13,14] are familiar traits often considered to affect physics learning but that would not normally be entirely subsumed under "preparation."

A key sentence of the paper comes in the discussion: "We initially expected that it would be differences in what high school physics courses were taken, but we analyzed that for HSWC [the highly selective west-coast university], and we found that all demographic groups at this institution had the same distribution of taking AP physics, regular high school physics, and no physics, even though the groups had different average CI prescores and math SAT or ACT scores" [25]. Thus course work, the main conventional component of preparation, was at least roughly measured in one college and was nominally the same for the different groups. That might suggest that preparation is *not* the key causal variable that differs. Other factors for which the precourse tests also serve as crude proxies have no such measured indicators of being matched between the different groups.

Realistically, however, courses with the same name can be radically different in different U.S. schools, and those differences are likely to show major correlations with racial or ethnic differences. Therefore, although the results provide no evidence for the role of preparation as distinguished from other causes, they also do not provide evidence against its role in causing differences between those groups. For the most part, however, males and females go to the same schools, so that if they took the same nominal courses they took the same actual courses. Thus, the results provide evidence that it is not the most obvious element of preparation, course work, that accounts for the male-female differences in both precourse and college exams. If preparation is taken in a broader sense to include all sorts of pretest factors, including unequal treatment of the genders both in and out of class, it becomes impossible to distinguish from other causes using these sorts of data.

These relations may again be easier to see with the help of a diagram. The paper [25] includes several DAGs used for SEM calculations of the predictive role of math ACTs and CI tests for different demographic traits. Although the

paper acknowledges that such SEM analysis "does not test for causality," the actual SEMs used can reasonably support a causal interpretation since the time order is clear: demography first, then pretests, lastly course tests [25]. Figure 3 here shows a simplified version of that diagram, but expanded to include the possible causes about which the paper makes claims. For simplicity I include only gender as a demographic variable and do not disaggregate the pretests, whose details are irrelevant here and were described in the original paper [25]. The problem is not with inference within the part of the graph shown in the paper, but rather that the key causal conclusions concerned potential causes that are not included in that part.
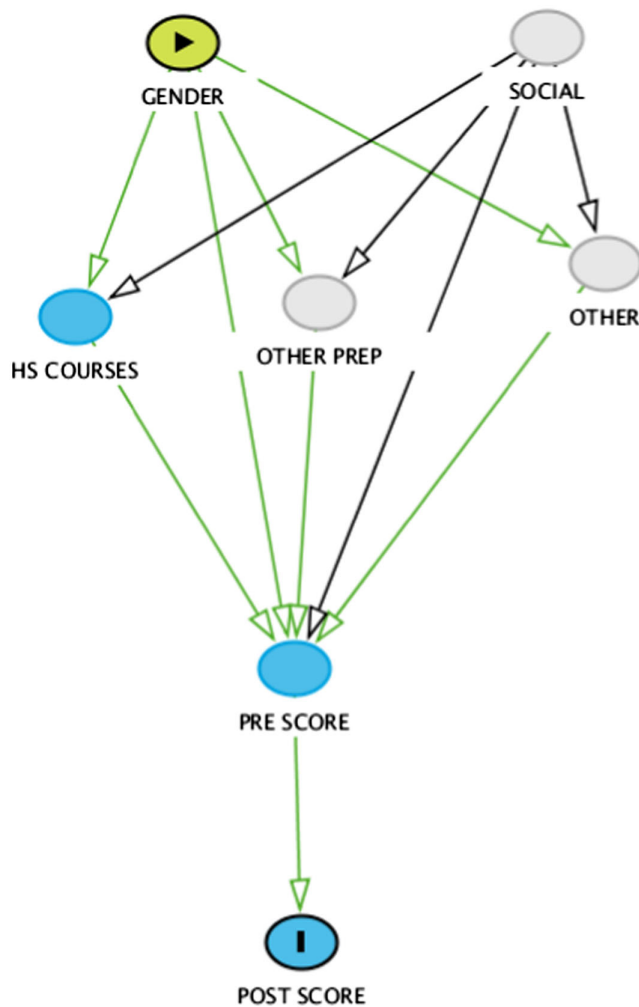

FIG. 3. A DAG schematically representing the causal issues addressed in Ref. [25]. Variables shown by lightly filled ovals are unmeasured. The combined effects of social variables and gender are presumably strongly nonlinear, i.e., the effects of gender depend on social context including bias. Since it happens that in this sample the variable HS courses is approximately independent of gender [25], the arrow between them could be erased. The observational result reported is that no arrow is needed from gender to postscore. Other demographic variables have similar diagrams but with poorly measured HS courses.

The central observational result is that the sum of the contributions of any paths from gender to postscores that bypass prescores is negligible. That result is unrelated to the causal claim about the differential effects of hypothetical changes in teaching methods [25], for which there are no relevant variables in the graph or the data. The causal claims about preparation concern the paths from gender to prescores, another region of the graph for which no relevant data are given. The only reported mediator on that path is high-school coursework, but since the effect of gender on it within this limited sample is said to be about zero [25], gender differences give no information on the effect of this major type of preparation on downstream variables. Thus, the data were consistent with erasing the arrow from gender to HS courses, if we ignore their effects on entry into this select cohort. All the other features are unmeasured. These include direct effects of gender on scores, and effects mediated by other forms of preparation or other experiences. All of these are moderated by the effects of the gender-asymmetric social environment, including biased treatment. Inclusion of the pretest variable tells us that some of these unmeasured variables have fairly long-lasting effects, but does not identify which variables.

The paper offers some options for specific policy recommendations. One broad option is to "provide more resources" [25], which presumably would help regardless of the specific causes of the individual differences and is not based on the study's data. Another option seems quite compatible with their interpretation of their data: "adding courses to the sequence to provide a greater range of students the opportunity to start with a course matched to their preparation" [25]. That policy has long since been implemented in some universities, e.g., my own, and it would be valuable to have studies of how well it works.

The primary option mentioned, however, is "…adjusting the course level to better match the preparation of the less prepared students…" This option is the one that the senior author amplified in his subsequent editorial based on the paper, advocating "changing the coverage and pace of some intro courses so they are optimized for the third of the distribution with the least preparation…" [26]. That policy recommendation seems less consistent with the preparation interpretation than with rejected causal interpretations. It may be a good way to modify some physics courses, but the data reported in the paper have no implications for that causal claim.

V. DISCUSSION

We have seen that several papers ignore the variety of causal possibilities consistent with a given dataset. This problem includes two familiar issues in causal inference: that correlations do not give the direction of causation and that they can be altered by confounding variables. We have also seen policy recommendations that are at best loosely connected with the causal interpretations.

Ideally policy recommendations should be based on reliable estimates of consequences. When the recommendations are instead intuitive guesses unconnected to our data, we should be open about that. The difficulties in drawing reliable causal conclusions from observational data in fields such as education should be taken as a reason to pay even closer attention to the relation between models and data than we do in physics, rather than as a reason to relax that attention. That task requires familiarity with modern causal inference methods.

The papers discussed here were chosen only to illustrate some of the pitfalls on the path from data to policy via causal inference. That it was not hard to find such papers from leading authors in a leading journal indicates that there is room for improvement in the editorial process. Adding some experts in causal inference to the editorial board of this journal (and others) might be helpful. One colleague suggested that perhaps a checkbox for referees indicating whether their reviews included evaluating causal inference would alert editors to the need for that further step without sacrificing the current evaluations of other aspects of the papers.

It might also help to have a causal reasoning primer specifically for PER, similar to those mentioned for other areas [3–9]. Its authors should include at least one with domain-specific PER knowledge and one with solid grounding in causal inference, both of which I lack. One referee suggested that a "healthy conversation" on causality would be good for the field.

Since I have suggested possible treatments it is important to also suggest how their effects might be evaluated. If an editorial change is made abruptly, perhaps a regression discontinuity analysis [27] could indicate whether it succeeded for the journal, at least if some reliable outcome measure were found. It would then be interesting to look for violations of the stable unit treatment value assumption [28], i.e., whether the level of causal reasoning was changed in just one journal, or perhaps raised throughout the field, or whether papers with problematic causal reasoning just moved to different journals. The effects of a primer or of some organized conversations would probably be more gradual and thus somewhat harder to disentangle from confounders.

In the meantime, readers who wish to do quick checks of whether reported observations are likely to be causal might start with the traditional informal criteria developed by Hill for public health questions [29]. Perhaps the most compact yet technically sophisticated modern account relevant for PER practitioners is also from the public health field [4], which shares many issues with educational research [9]. As a reader, I have found it useful to sketch a causal graph and interpret words in a paper as statements about links that are on the graph or that are omitted. A fuller description of useful methods and of the many pitfalls (reverse causation, confounding variables, improper controls eliminating causal paths, various forms of selection bias, lack of generalizability, etc.) can be obtained from many texts, e.g., Refs. [1,2]. That is far beyond the scope of this brief paper, which is intended only to call attention to the issues.

ACKNOWLEDGMENTS

[1] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics—A Primer* (Wiley, Chichester, UK, 2016).

[2] M. A. Hernán and J. M. Robins, *Causal Inference: What If* (Chapman & Hall/CRC, Boca Raton, 2020).

[3] C. Glymour, K. Zhang, and P. Spirtes, Review of causal discovery methods based on graphical models, Front. Genetics **10**, 524 (2019).

[4] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet, Causal inference in public health, Annu. Rev. Public Health **34**, 61 (2013).

[5] H. R. Varian, Causal inference in economics and marketing, Proc. Natl. Acad. Sci. U.S.A. **113**, 7310 (2016).

[6] E. M. Foster, Causal inference and developmental psychology, Dev. Psychol. **46**, 1454 (2010).

[7] M. Gangl, Causal inference in sociological research, Annu. Rev. Sociol. **36**, 21 (2010).

[8] L. Keele, The statistics of causal inference: A view from political methodology, Political Anal. **23**, 313 (2015).

[9] R. J. Murnane and J. B. Willett, *Methods Matter: Improving Causal Inference in Educational and Social Science Research* (Oxford University Press, Oxford, UK, 2011).

[10] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 4th ed. (Guilford Publications, New York, 2015).

[11] K. A. Bollen and J. Pearl, in *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan (Springer, Dordrecht, 2013), Chap. 15, p. 301.

[12] T. Richardson, Markov properties for acyclic directed mixed graphs, Scandanavian J. Stat. **30**, 145 (2003).

[13] Z. Y. Kalender, E. Marshman, C. D. Schunn, T. J. Nokes-Malach, and C. Singh, Why female science, technology, engineering, and mathematics majors do not identify with

physics: They do not think others see them that way, Phys. Rev. Phys. Educ. Res. **15,** 020148 (2019).

[14] R. M. Lock, Z. Hazari, and G. Potvin, Impact of out-of-class science and engineering activities on physics identity and career intentions, Phys. Rev. Phys. Educ. Res. **15,** 020137 (2019).

[15] J. Robins, A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, Math. Modelling **7,** 1393 (1986).

[16] M. A. Hernán, S. Hernández-Díaz, and J. M. Robins, A structural approach to selection bias, Epidemiology **15,** 615 (2004).

[17] K. R. Binning, N. Kaufmann, E. M. McGreevy, O. Fotuhi, S. Chen, E. Marshman, Z. Yasemin Kalender, L. Limeri, L. Betancur, and C. Singh, Changing social contexts to foster equity in college science courses: An ecological-belonging intervention, Psychol. Sci. **31,** 1059 (2020).

[18] J. D. Cribbs, Z. Hazari, G. Sonnert, and P. M. Sadler, Establishing an explanatory model for mathematics identity, Child Development **86,** 1048 (2015).

[19] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control **19,** 716 (1974).

[20] S. Greenland, Quantifying biases in causal models: Classical confounding vs collider-stratification bias, Epidemiology **14,** 300 (2003).

[21] S. Hidi and K. A. Renninger, The four-phase model of interest development, Educ. Psychol. **41,** 111 (2006).

[22] Z. Hazari, G. Sonnert, P. M. Sadler, and M.-C. Shanahan, Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study, J. Res. Sci. Teach. **47,** 978 (2010).

[23] Z. Hazari, G. Potvin, R. M. Lock, F. Lung, G. Sonnert, and P. M. Sadler, Factors that affect the physical science career interest of female students: Testing five common hypotheses, Phys. Rev. ST Phys. Educ. Res. **9,** 020115 (2013).

[24] C. S. Hulleman and J. M. Harackiewicz, Promoting interest and performance in high school science classes, Science **326,** 1410 (2009).

[25] S. Salehi, E. Burkholder, G. P. Lepage, Steven Pollock, and C. Wieman, Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics, Phys. Rev. Phys. Educ. Res. **15,** 020114 (2019).

[26] C. Wieman, It's Not "Talent," it's "Privilege", APS News **29,** No. 9, 8 (2020).

[27] D. S. Lee and T. Lemieux, Regression discontinuity designs in economics, J. Econ. Lit. **48,** 281 (2010).

[28] S. Schwartz, N. M. Gatto, and U. B. Campbell, Extending the sufficient component cause model to describe the stable unit treatment value assumption, Epidemiol. Perspect. Innov. **9,** 1 (2012), https://pubmed.ncbi.nlm.nih.gov/22472125/.

[29] A. B. Hill, The environment and disease: Association or causation?, Proceedings of the Royal Society of Medicine **58,** 295 (1965).