Running head: CAUSAL COMMENTS

Commentary on Causal Prescriptive Statements

Arthur C. Graesser and Xiangen Hu

University of Memphis

Send correspondence to:

Art Graesser
Department of Psychology & Institute for Intelligent Systems
202 Psychology Building
400 Innovation Drive
University of Memphis
Memphis, TN 38152
901-678-4857
901-678-2579 (fax)
a-graesser@memphis.edu

Educational Psychology Review, submitted

Commentary on Causal Prescriptive Statements

Causal prescriptive statements are necessary in the social sciences whenever there is a mission to help individuals, groups, or organizations improve. Researchers inquire whether some variable or intervention A causes an improvement in some mental, emotional, or behavioural variable B. If they are satisfied that A causes B, then they can take steps to manipulate A in the real world and thereby help people by enhancing B. It is this utility in helping people that makes it imperative to test causal prescriptive statements with rigorous methodology and to rule out extraneous third variables and confounding variables as being the true cause instead of variable A. Aside from the utility of helping people, tests of causal prescriptive statements are important for advancing scientific theories of the mechanisms that underlie cognition, emotion, and behaviour. Causal statements are not essential for advancing all sciences. Astronomy and some areas of physics are essentially descriptive so researchers are not obsessed with questions of causality. Nevertheless, rigorous tests of causal statements are always a welcome addition to a science.

Consider an example with no relevance to educational psychology. Most of us have had the experience of our automobile engine breaking down, hopefully not on a dangerous highway. Who would be the best person to help us in this situation? The media junkie who reads *Consumer Reports* on hundreds of automobiles with respect to dozens of performance measures? Or the mechanic who adjusts fuel, mechanical, and electrical controls to explore what gets the engine going? The answer is perfectly obvious in a mechanistic world. We want someone who can manipulate one or more of the controls that causally results in the engine changing from off to running. This can be accomplished in a niche that is carefully

understood and engineered to the point where all of the relevant variables are pretty much known and unlikely to change over time. The repertoire of relevant parameters is complete, adequately specified, and invariant within a well-defined environmental niche.

Consider next an example that is very relevant to educational psychology. Some reading researchers believe it is important to tailor training to the individual needs of the students. That is, reading intervention 1 is best for one group of readers and reading intervention 2 is best for another group of readers. The tailored intervention is expected to help students read better than a uniform, judiciously selected, scripted intervention that all students receive. Other reading researchers (names intentionally withheld) do not agree. They believe that students learn how to read best by a well crafted scripted intervention that is faithfully delivered. So the question is whether tailoring the intervention causes an improvement in reading skills in students. To test this causal prescriptive statement, the researchers randomly assign students to either a tailored intervention or a scripted control condition and then measure reading achievement scores months or years later. Tailoring is supported if the reading scores in the tailored intervention group significantly exceed the scores in the scripted intervention group. Random assignment with a large sample rules out extraneous variables as being responsible for the advantage of tailored over scripted control. The possibility of confounding variables being responsible also needs to be considered, as will be discussed later. However, for the most part, this randomized control trial is the methodological gold standard for testing such a causal prescriptive statement.

The reading example differs from the automobile example in many ways. One aspect is timing. The mechanic will quickly discover whether a control causes the engine to turn on so

a large number of actions can be tested out in a small amount of time. In contrast, it may take years for the reading researcher to find out whether the tailored intervention has a significant improvement over the scripted control condition. Another aspect is whether the set of relevant variables is closed and complete versus open and incomplete. The automobile engine was designed by a group of engineers to handle a finite, closed set of operational variables within the well-specified niche. In contrast, the set of relevant variables to reading researchers is open and indeterminate in an uncertain landscape of niches. Yet another aspect is whether the system is likely to vary over time and history. The socio-cultural history of people changes dynamically over time whereas the physical environment of automobiles is essentially constant. These differences have enormous implications on the difficulty of testing causal prescriptive statements in reading research.

The papers in this special issue edited by Kulikowich and Sperling provide different perspectives on testing causal prescriptive statements. The articles explore the status of causal prescriptive statements (CPSs) from the perspective of qualitative methods (Nolen & Talbert), logical and statistical inference (Sun & Pan), structural equation modelling (Martin), longitudinal models that compare a treatment to control (O'Connell & Gray), randomized control trials (Marley & Levin), propensity scoring (Bai), psychometric assessment with grounding in cognitive models (Brown & Wilson). Each of these perspectives has a rich intellectual history with assorted virtues and trappings. They all adopt the foundations of research design and statistics that are widely accepted in the social sciences (Shadish, Cook & Campbell, 2002), but they also offer distinctive methodological slants that may not be universally shared. However, they do share the assumption that CPS's are extremely important

to assess and acknowledge the ubiquitous problem that researchers make causal claims when the supporting data is either correlational or descriptive (Robinson, Levin, Thomas, Pituch, & Vaughn, 2007).

This commentary has a core take-home message: It is extremely difficult to validate causal prescriptive statements empirically, possibly more difficult than what is portrayed in this collection of 7 articles. Defence of a CPS must therefore be advanced with an air of humility because it is easy to kill it with a plausible counterargument, critique, or data set. Nevertheless, that should not discourage researchers from trying.

The analysis of causality that we have adopted in the social sciences has four criteria: Temporality, operativity, necessity, and sufficiency (TONS as an acronym).

Temporality. Variable A must precede the effect on variable B in time. This assumption is adopted by everyone, including those contributors to this journal issue who offered definitions of causality.

Operativity. The time span of variable A must be intact, in whole or part, when observing variable B. This criterion was not mentioned by any of the contributors, but it is clearly one that should be considered. If the intervention A has dampened to the point of no longer operating, than there should be no expected change in B. Operativity is related to the contiguity criterion that A and B must be close to each other temporally. However, sometimes an intervention A has a delayed impact on variable B in the social sciences, so the dynamics of operativity bears scrutiny.

Necessity and sufficiency. Necessity or sufficiency must be established, but both are not required. The necessity criterion must satisfy the counterfactual test: If A does not occur,

then B does not occur. Necessity is rarely established in the social sciences, and many phenomena in the natural sciences, because most psychological events are caused by one or more of several alternative sufficient causes. Sufficiency receives the most attention. We view sufficiency from a systems perspective rather than a logical definition: If A occurs and the background circumstances continue as usual, then B occurs (Mackie, 1975). This is a weaker sense of sufficiency than would be adopted in Sun and Pan's article in this issue.

The contributors to this special issues offer additional criteria to be included in CPSs. Some try to quantify the likelihood, effect size, or statistical reliability of A causing B. This provides a more precise specification of the sufficiency criterion. Some offer a relativistic account by showing that A's causing B is more robust than another condition C's causing B. Most researchers spend considerable effort making sure that a third variable T or an unexpected confounding variable U (unintentionally triggered by A) is not responsible for B rather than intended manipulated variable A per se. The need for replications in different contexts and populations helps researchers evaluate the contributions of T and U, although not completely. So does an analysis of the data from the standpoint of theoretical models.

All of these constraints make it abundantly clear that is difficult to establish a prescriptive statement that A causes B. The strongest case can be made when researchers conduct randomized control trials and replicate the results in many states, schools, classrooms, teachers, and students. That is obviously an expensive and time-consuming effort. Moreover, the data are often not kind, in the sense that sometimes effects are not replicated, sometimes the effect sizes are small in magnitude, and sometimes there are aptitude-treatment interactions

(that may or may not be examined). Unfortunately, some of the most robust CPS's are uninteresting theoretically or are impractical to scale up.

However, the world is even more complex in rigorous evaluations of CPSs. Consider some of the following challenges, even for expensive tests with the randomized control trials (RCT). In order to illustrate the claims more concretely, I will use examples in an area of research that we have conducted, namely intelligent tutoring systems with animated conversational agents. These computer agents help students learn by holding conversations in natural language and that respond adaptively to the students' cognitive and emotional states (D'Mello & Graesser, 2010; Graesser, Rus, D'Mello, & Jackson, 2008; Hu & Graesser, 2004; VanLehn et al., 2007).

(1) ***Indeterminacy challenge***. Everyone agrees that a large number of third variables (T) may be responsible for a correlational relationship between A and B. Analogously, a large number of unexpected confounding variables (U) may be responsible for a successful RCT that tests A causes B. For example, suppose that a researcher claims that a conversational agent helps students learn physics at a deeper level compared to reading a textbook for an equivalent amount of time (VanLehn et al., 2007). The researcher would be thrilled if the interactive conversational agent yielded significant learning gains in an RCT with a posttest measure that taps deep learning. Unfortunately, however, we would not know whether the improved comprehension stems from the conversational interactivity and cognitive learning, as intended by the developers, or whether the cause was instead an enhanced motivation of the students who experienced a novel agent environment, boredom from reading a dull physics book, or some other confounding variable that bears little relation to the intended theoretical

manipulation. Which is the most problematic: A potentially infinite number of third variables or a potentially infinite number of confounding variables?

(2) ***Interactions challenge***. Aptitude-treatment interactions can potentially cloud a test of A causes B, as Cronbach and Snow (1977) documented decades ago. Some students might be helped by an intervention but other students may be harmed, thereby cancelling each other out. For example, we have been investigating the impact of conversational agents who do or do not have an emotional sensitivity to the learner (D'Mello et al., 2010; Graesser et al., 2008). One recent study revealed an interesting interaction. A conversational agent that is emotionally supportive of students through language and facial expressions tends to help low knowledge students learn, but interferes with learning of high knowledge students compared to an agent that is not emotionally sensitive. The net effect is that emotional sensitivity has no overall impact on the entire set of students. However, it would be misleading to say that affect sensitivity has no impact in light of the significant aptitude-treatment interaction. This example illustrates an important challenge form variable interactions. There are many interactions that a researcher could test when considering potential covariates. Unfortunately, most educational researchers test very few, if any, such interactions.

(3) ***Combinatorial explosion challenge***. The combinatorial explosion of potential variables is virtually never investigated. We have often wanted to know what types of agents help particular types of students learn best (Graesser et al., 2008; Hu & Graesser, 2004). For example, do students learn best when they are matched with the agent on gender, ethnicity, and age? Or is there a more complex set of attributes that are more optimally aligned between student and agent. One could easily generate 20 likely attributes of the tutor agent and 20

attributes of the student in a test that makes an attempt to be thorough. If there are 40 variables and merely 2 values per variable, there are over a trillion combinations to test - far too many. As a consequence, researchers turn to their theories to narrow down the set of variables to a smaller manageable number. The use of theory is indeed quite sensible. However, it is not enough to stop the critic from claiming that the researcher has not ruled out a staggering number of confounding variables, covariates, and interactions, even in RCT's. It is rare to find a published study that seriously takes on the combinatorial explosion challenge.

(4) ***Dynamics challenge***. Background variables, people, interventions, and psychological effects all dynamically change over time, sometimes in complex ways (Spivey, 2007). It is easier for researchers to thing in terms of traits, constancies, and discrete separate events than to keep track of the dynamics even when the dynamics are well known. Fashions change. An animated agent that is cool and funny to students one month look nerdish and "yesterday" the following year. Emotions change. A conversational agent may seem novel and flashy the first half hour but ends up becoming stale and irritating on the 10th hour. Lifespan development changes. The conversational agents are interesting to the young generation but frivolous to the aging generation. The general challenge of dynamics is complicated by the fact that there are so many dynamical trajectories to consider: exponentially decreasing functions, attraction basins, oscillations, recurrent feedback loops, boomerang effects, hysteresis, and other patterns that are beyond the scope of this paper to define. All of these possibilities of course aggravate the combinatorial explosion challenge.

With this context in mind, it is time to turn to more specific comments on the 7 contributions in this special issue. We will identify one positive message and one negative

note on each. This balanced commentary will hopefully augment a rich set of papers that should guide researchers in their tests of CPS's. There will also be an ordering in papers addressed. We start with papers that primarily address descriptive research, and then go on to correlational methods, and finally to experimental designs that manipulate variables.

The Nolen and Talbert article emphasizes prescriptive statements with qualitative assertions. There is little or no attempt to deconstruct causal statements in this descriptive approach to prescriptive statements. These researchers identify a variety of qualitative methods that specify particular cases in rich detail, including phenomenology, narrative inquiry, and case studies. The value of this research is that it specifies individual cases in rich detail, with transparency on how researchers apply methods in generating such content. The careful detail in documenting cases is essential if we want to take stock of potential measures of background circumstances, third variables, and potential confounding variables. However, there are drawbacks to this approach. There are a staggering number of attributes to document so researchers will have little patience to wade through a sea of descriptions without theoretical guidance and sensitivity to priorities of importance. Such considerations of focus clash with a neutral documentation of the case attributes. Cherry-picking the cases to explore runs the risk of biasing the sample of observations. A systematic approach to sampling and analysis of qualitative data is a virtue that is pursued by fields such as corpus analysis, computational linguistics, and discourse processing (Graesser, Gernsbacher, & Goldman, 2003). These fields are not the same as those that select cases and attributes conveniently in order to advance a research agenda or make a rhetorical point. Cases that are used to illustrate a point are rhetoric, not science, whereas cases that are systematically sampled and

decomposed are indeed in the arena of science. We live in an era when it is important to perform qualitative analyses of quantitative data, and quantitative analyses of qualitative data. Pure qualitative data and pure quantitative data sets are arguably limited.

The Martin contribution shows how structural equation modelling (SEM) can be used to test causal statements from correlational data. The pattern of cross-panel correlations at two points in time help the researcher converge on what might be causing what. Latent growth modelling helps the researchers document complex trajectories over time in longitudinal studies. A systematic SEM can track moderator and mediator variables in addition manipulated and outcome variables. The SEM approach with longitudinal or cross panel designs are very rigorous attempts to extract causality from a bewildering sea of correlations. The down size of this approach is that the set of variables fluctuates from study to study, including interactions between/among predictor variables. The use of SEM in experimental studies with manipulated variables is also less prominent than is hoped. The typical routine is for researchers to administer tests with a large set of variables on two occasions and to report the trends that rise to the surface. This approach requires replications on new samples and schools so that stable structures can be detected and confirmed. Another problem with this approach is that true causal relationships may not emerge because of the sea of covariates and confounding variables that are too often invisible to the researcher.

The Bain article applies propensity score analysis to equilibrate characteristics of treatment versus comparison groups in observational studies. A propensity score analysis matches a treatment versus comparison condition on a set of covariates in an attempt to rule out third variables. This is an important step, particularly if the sample size is modest.

However, there is no guarantee that all of the relevant third variables are measured, that confounding variables are sufficiently addressed, and that an adequate number of interactions are tested.

The O'Connell and Gray article extols the virtues of longitudinal models that have a dichotomous contrast between a treatment and control condition. This is a major enhancement over normal RCT's because the researcher can trace predominant trends over time, subclasses of trajectories, and potentially personal trajectories. One limitation of this approach in practice (as opposed to what is theoretically possible) is that researchers have entertained only a small number of functions and small number of points in time. Instead of considering growth or complex functions that fit data points to 3-6 points in time, imagine measures collected at dozens, hundreds, or thousands of points in time, as is often done in the fields of educational data mining (Baker & Yacef, 2010) and intelligent tutoring systems (D'Mello & Graesser, 2010; Ritter, Anderson, Koedinger, & Corbett, 2007). This would substantially increase the family of trajectories to explore and rigorously test.

Marley and Levin articulate the virtues of the RCT, the gold standard of testing causal prescriptive statements. They identify the typical tradeoffs between the establishment of internal and external validity. Their proposed CARE model documents the practical stages in research of starting with the documentation of observations and correlations, then on to replications, followed by testing causal relationships in RCT's and conducting additional studies to rule out alternative explanations. This approach has been adopted by the Institute of Education Sciences in addition to the field of medicine. This mainstream methodology will perhaps always be adopted in education. Its major limitation is that it underestimates the

seriousness of the four challenges mentioned earlier: indeterminacy, interactions, combinatorial explosion, and dynamics. RCT will never go the distance in scaling up to accommodate open-ended complex mechanisms when there is a finite amount of time, expense, and sample size. The alternative is to have a methodology that collects a richer amount of data from each student (or other class of observations), as well as more complex mathematical models. However, such a methodology would unfortunately force researchers to weaken claims on causal prescriptive statements.

The Sun and Pan article appropriately emphasizes the importance of replication, meta-analyses over several studies, and more advanced Bayesian approaches to testing CPS's. There is also some practical guidance in writing Discussion sections. All of these recommendations are technically on the mark. The major drawback is that most researchers are not currently trained to conduct the complex Bayesian analyses and there are not enough trained quantitative methodologists to fill the gap. This research training gap was acknowledged Kulikowich and Sperling's introduction. The gap can perhaps be mitigated by practical guidance in documents or software that can be understood by researchers from diverse backgrounds. However, the value added by the more sophisticated quantitative methods over traditional methods needs careful scrutiny. What is fundamentally gained when the sophisticated quantitative analyses, which can be conducted by only 100 experts in the field, yield only a 2% gain in accuracy over simpler quantitative analyses that can be competently conducted by 100,000 researchers?

Brown and Wilson must be applauded when they insist on a model of cognition to guide the construction of assessments. One of the major breakthroughs in psychometrics during the

last decade has been theory and conceptually guided assessment measures, as articulated under the banner of Evidence Centered Design (Mislevy, 2007). Instead of reliability and atheoretical psychometrics ruling the roost of measures, there are links to psychological and pedagogical theories to guide item design. This guidance is perfectly aligned with theory and one is reminded of the cliché that there is nothing more practical as a good theory. Brown and Wilson propose learning progression models as an example application of evidence centered design. Unfortunately, their example is too simple and unidimensional. Modern cognitive models are extremely complex. There are nonlinearities, interactions, complex feedback mechanisms and other features that outstrip the simple ordering of a unidimensional scale. Until advanced psychometrics emancipates itself from unidimensional psychometrics, the field will never meet the constraints of modern dynamical systems models that are ubiquitous in cognitive science (Spivey, 2007).

In closing, it is important to return to the main intended message of this commentary. Simply put, it is extremely difficult to confirm causal prescriptive statements in the social sciences. The articles in this journal uniformly support this claim. There are two implications to this conclusion. First, we need to be very explicit and strategic in scientific research that is targeted to CPS's because the research is complex, time-consuming, and expensive. Second, we need to take stock of the methods outside of the realm of CPSs that contribute substantially to understanding phenomena. Much can be learned by description, as every physicist knows. However, description alone will never would have put us on the moon, cured infections, or advanced educational research. CPS's are here to stay, even though we recognize the obstinate difficulties in validating them.

References

Baker, R., Yacef, K. (2010). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1,* 3– 17.

Cronbach, L. & Snow, R. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

D'Mello, S., & Graesser, A.C. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction, 20*, 187.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. (Eds.). (2003). *Handbook of discourse processes*. Mahwah, NJ: Erlbaum.

Graesser, A. C., Rus., V., D'Mello, S., & Jackson, G. T. (2008). AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), *Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning* (pp. 95–125). Information Age Publishing.

Hu, X., & Graesser, A. C. (2004). Human Use Regulatory Affairs Advisor (HURAA): Learning about research ethics with intelligent learning modules. *Behavior Research Methods*, *Instruments*, *and Computers*, *36*, 241–249.

Mackie, J.L. (1975). *The cement of the universe: A study in causation.* Oxford: Clarendon Press.

Mislevy, R.J. (2007). Validity by design. *Educational Researcher, 36*, 463-469.

Ritter, S., Anderson, J. R., Koedinger, K. R., Corbett, A. (2007) Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*, 249-255.

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. R. (2007). The incidence of "causal" statements in teaching and learning research journals. *American Educational Research Journal, 44*, 400–413.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Spivey, J.M. (2007). *The continuity of mind*. Oxford: Oxford University Press.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.