International Journal of Aviation Research

Volume 13 | Issue 01 Peer-Reviewed Article

Structural Equation Modeling in Aviation: Is It Too Good to Be True?

David Trafimow
New Mexico State University

It is difficult for aviation researchers to perform complex experiments to demonstrate various types of causation such as mediation, moderation, and so on. Therefore, structural equation modelling (SEM) applied to correlational studies is increasingly becoming de rigueur in the aviation area. To engage in SEM, it is merely necessary to measure all the constructs of interest, and the SEM program provides the path coefficients, significance tests, goodness of fit indices, and all the rest. It is extremely convenient to be able to replace difficult-to-perform experiments with relatively easy-to-perform SEM studies. Therefore, it is not surprising that SEM is becoming increasingly favored. However, the easiness is suspicious. The present article examines carefully whether SEM soundly fulfills its promise to provide strong evidence of causation or of the extent of the causation, as indexed by path coefficients.

Recommended Citation:
Trafimow, D (2021). Structural equation modeling in aviation: Is it too good to be true? *International Journal of Aviation Research*, *13*(01), 124-147.

Structural equation modeling (SEM), applied to correlational studies to draw causal conclusions, is fast becoming the most favored paradigm in aviation research (e.g., Fussell and Truong, 2021; Jenatabadi, & Ismail, 2014; RaviKumar, Ramanathan, & Porkodi, 2012; Singh, Vaibhav, & Sharma, 2021; Winter, Crouse, & Rice, 2021; Winter, Keebler, Lamb, Simonson, Thomas, & Rice, 2021). Nor is this a unique trend; SEM is increasingly becoming favored in many fields as diverse as management, marketing, several areas of medicine, and several areas of psychology. It is worthwhile to consider the reasons for the trend, with a special focus on aviation, as this journal is the "*International Journal of Aviation Research.*" The subsequent paragraphs discuss that (a) aviation researchers wish to draw causal conclusions and (b) it is easier to perform correlational than experimental research.

Imagine a universe where only true experiments, with manipulation of the putative causal construct (e.g., Campbell, Stanley, & Gage, 1963), were legal. In this restricted universe, suppose a researcher wishes to demonstrate that *A* causes *B*. The researcher likely would experimentally manipulate *A* and hope for an effect on *B*, a reasonably straightforward project. But consider another researcher who wishes to demonstrate that *A* causes *B* causes *C*, or to use the loaded word, *B mediates* between *A* and *C*. Fulfilling our second researcher's desires is not straightforward and might require multiple experiments. The researcher might manipulate *A* and obtain effects on *B* and *C*, but this would not demonstrate that the effect of *A* on *C* is through *B*. A potential second experiment would be to again manipulate *A*, with the addition of an orthogonal manipulation designed to fix *B* by driving it to a floor or ceiling. The hope would be that manipulating *A* influences *C* when *B* is not fixed at a floor or ceiling, and thereby allowed to vary, but manipulating *A* does not influence *C* when *B* is fixed at a floor or ceiling. Although this strategy has been used successfully (e.g., Trafimow et al., 2005), it is not straightforward. In addition to having to figure out how to manipulate *A*, the researcher also must solve the difficult puzzle of how to perform an orthogonal manipulation that fixes *B* or does not fix *B* at some level. The inclusion of additional constructs in the causal chain increases the difficulties.

However, we do not live in a restricted universe. The typical answer to the problem is what Wright (1934) termed the *method of path coefficients* that has evolved into *structural equation modeling* (SEM) (e.g., Bollen & Pearl, 2013; Byrne, 2013; Jöreskog & Sörbom, 1979; Heise, 1975; Kenny, 1979; Pearl, 2011). Rather than performing difficult, and perhaps unfeasible, experiments, the researcher can simply measure all the variables of interest, include them in a model, and let the SEM program supply the path coefficients. The SEM program also can supply significance tests of all path coefficients, goodness of fit indices, and an overall significance test of the whole model. If the path coefficients, *p* values, and fit indices pan out, the researcher can claim to have demonstrated the hypothesized causal model. Consequently, it is unsurprising that researchers, across many sciences, have moved, or are moving in this direction. As stated previously, the field of aviation research is one of those sciences (e.g., Fussell and Truong, 2021; Jenatabadi, & Ismail, 2014; RaviKumar, Ramanathan, & Porkodi, 2012; Singh, Vaibhav, & Sharma, 2021; Winter, Crouse, & Rice, 2021; Winter, Keebler, Lamb, Simonson, Thomas, & Rice, 2021).

Contrasting both foregoing paragraphs suggests that aviation researchers would have to be crazy to perform experiments to test hypotheses involving mediation, moderation, mediated moderation, moderated mediation, and so on. As we have seen, experimental work to test hypotheses more complex than those involving one putative causal construct is difficult to conduct, whereas SEM handles complexity with ease. When researchers use SEM, more variables simply mean the researcher needs to include more measures, but this is relatively easy compared to attempting to address causal complexity via true experiments, with the necessity to manipulate crucial hypothesized constructs and perform orthogonal manipulations to address potential mediating or moderating constructs. SEM would seem too good to be true, except that many authorities have supported its soundness (e.g., Bollen & Pearl, 2013; Byrne, 2013; Jöreskog & Sörbom, 1979; Heise, 1975; Kenny, 1979; Pearl, 2011). Those researchers who fail to get onboard the SEM bandwagon are making their lives immensely more difficult than need be and failing to use one of the most powerful analytic paradigms ever invented.

Or is SEM too good to be true, after all? There is preliminary reason for suspicion. Trafimow (2015) tested two extremely wrong models of planetary mass, velocity, and momentum, using the method of path coefficients. One model was that planetary mass causes planetary velocity causes planetary momentum whereas the other model was that planetary velocity causes planetary mass causes planetary momentum. Although the results of the path analyses disconfirmed the former model, the latter one was strongly confirmed. As both models are blatantly false, there is reason for doubting that this method is as definitive as it is touted to be. But why?

The Causation Issue

Whether researchers admit it straightforwardly or not, there can be little doubt that they use SEM to draw causal conclusions from correlational data. This fact is rendered obvious by arrows in diagrams. What could the arrows represent other than causation? And if that were not sufficient, there is inevitably one or more recommendations for policy change, intervention, or others that would make no sense without assuming causation. Do applications of SEM to correlational studies justify causal conclusions?

To answer, let us commence with a reminder of introductory psychology, marketing, economics, etc., where teachers tell students that correlation does not mean causation. If *A* and *B* correlate, it could be that *A* causes *B*, *B* causes *A*, or some other variable, or set of variables, causes both. Because of the ease of generating alternative causal explanations, the correlation, by itself, fails to provide a strong case for any single causal hypothesis.

In contrast, when there are multiple variables at play, and a researcher uses SEM, alternative explanations come to mind less easily. Thus, it is possible to argue that SEM is superior to single correlations because single correlations can be explained away easily

whereas SEM findings cannot. But a counterargument is that with increased model complexity, it is more difficult to posit alternative explanations because the increased model complexity forces increased cognitive complexity too. Thus, the difficulty in positing alternative explanations is not because they are not there, but because of human processing capacity limitations. To render salient the possibilities, Kline (2015) demonstrated 18 plausible models for three-variable cross-sectional designs when there is no strong independent rationale for directionality specification (also see Grice et al., 2015; Tate, 2015; Thoemmes, 2015).

To the complexity issue, Trafimow (2017; also see Saylors & Trafimow, 2021) demonstrated that increased complexity decreases the probability that the causal model is correct. To illustrate quickly, imagine only two variables, where there is a single correlation coefficient. As we have seen, there are alternative explanations, but there might at least be a reasonable probability that the correlation is for the correct reason (the reason hoped for by the experimenter). But now let us imagine three variables and three correlation coefficients underlying SEM. We might ask about the probability that all three are for the correct reason. Suppose we generously assume that each correlation coefficient has a probability of 0.7 of being for the right reason. Well, then, in the case of a single correlation coefficient, the probability of the simple two-variable model being correct would be 0.7. But for the three-variable model, the model probability would 0.7 x 0.7 x 0.7 = 0.343 (unless the correlations are dependent on each other, but even then, the basic arithmetic still holds conceptually, Trafimow, 2017). In the case of a four-variable model, there are 6 underlying correlation coefficients, and now the model probability decreases to 0.2401. For a five-variable model, there are 10 underling correlation coefficients, and the model probability decreases to 0.16807. And so on. Although many find this surprising, it should not be. As the model becomes increasingly complex, the probability increases that there is something wrong somewhere.

A way to address the truism that correlation does not mean causation, and perhaps it could even somewhat address the complexity issue (Kline, 2015; Trafimow, 2017), is to insist that there can be independent reasons for assuming causation. And there are at least two directions in which this can go. One direction is to focus on using the causal pathways for which there is independent support for causation to test additional potential causal pathways. A second direction is to focus on using SEM to find the extent of the causation by obtaining path coefficients. The following two subsections address each, in turn.

Using Assumed Pathways to Test for Speculative Ones

Let us commence by considering that in SEM contexts, causation is defined as having nonzero path coefficients. There is little point in arguing about whether the definition is correct because definitions are always correct by fiat. The question is whether the definition is useful, not whether the definition is correct. In this spirit, consider too that there is an infinitude of possible values that path coefficients can have, so it is unlikely for a path coefficient to have a value exactly equal to zero. And even if the theoretical value were zero, imperfections in the measures of the variables, in the

execution of the study, and so on, would almost certainly guarantee a value not exactly equal to zero in any single study (even ignoring the issue of sampling error). There is no escaping that the SEM definition of causation renders tantamount to certain that causation is always there in the form of a nonzero path coefficient!

With the definitional issue clear, we are now ready to address the use of SEM to use pathways assumed true to test whether additional pathways are true. It should be obvious that because all non-zero pathways indicate causation, by definitional fiat, the new hypothesized pathways can be assumed to be there too. In fact, all pathways should be assumed there, even without conducting a SEM; after all, how likely is it that a path coefficient will exactly equal zero?

A potential way to push back on this argument engages null hypothesis significance testing (NHST). The idea would be that only those pathways with statistically significant path coefficients are to be taken as indicating causation, with those that are not statistically significant taken as indicating a lack of causation. However, there are multiple problems with this criterion. As a preliminary, even according to party line NHST thinking, significance tests can only be used to show that an effect is there but cannot be used to show that an effect is not there. Thus, there remains no way to conclude that there is a lack of causation, even if the *p* value comes back, say, at the value of 0.99.

More important, researchers fail to fully appreciate that the *p* values featured in NHST are not conditioned on the null hypothesis, as is typically taught in statistics courses, but rather are conditioned on a full statistical model. It is crucial to distinguish between hypotheses and the models in which they are embedded. The model includes not just the null hypothesis, but a large set of additional assumptions too. There are so many additional assumptions that Bradley and Brand (2016) and Trafimow (2019a) proposed taxonomies to organize them. For example, a ubiquitous assumption is that the researcher has sampled randomly and independently from the population (Berk and Freedman, 2003; Hirschauer, 2020). In the tradition of Greenland (2019), we can denote the set of additional assumptions as follows: $A = \{a_1, a_2, \ldots, a_n\}$. Thus, as the model *M* includes the null hypothesis (or test hypothesis) *H*, and the set of additional assumptions *A*, we have a concise equation: $M = H + A$. Moreover, *A* is always wrong. For instance, there is no research in the aviation field where participants were randomly and independently sampled from a defined population. As $M = H + A$, and *A* is always false, it follows deductively that *M* is false too. And aviation researchers ought to face this squarely.

Because *p* values are conditioned on *M*, as opposed to *H*, they do not index the incompatibility of the evidence with the *hypothesis*; they only index the incompatibility of the evidence with the *model*. But as the model is known wrong anyway, the *p* value provides little useful information about it. Put in the form of a pointed question: Why should researchers work hard to obtain evidence against a model already known to be wrong?

A potential answer is that the model might be close to right, though not exactly right, and if the model is close enough to right, it might still be useful. I agree that models

can be useful, though wrong. If one uses a model to infer a population parameter from a sample statistic, the estimate will be false, but it might be close enough to true to be useful. But the problem is that in SEM, we are not using the statistical model for estimation. We are using it to make dichotomous decisions about causation being there or not being there. As the old saying goes, there is no such thing as being slightly pregnant.

Further appreciating the distinction between hypotheses and the models in which they are embedded implies an additional problem. Although *p* values provide information about the incompatibility of the evidence with the models under which they are conditioned, they do not provide information about the incompatibility of the evidence with hypotheses embedded in those models. A wee *p* value could be due to the additional assumptions being wrong, or to the hypothesis being wrong too. There is no way to know. To illustrate the issue of null hypotheses being embedded in larger models, consider the analogy of an ugly art museum that contains paintings; that the museum is ugly fails to provide a sound reason for concluding that the paintings within are ugly too.

Nor do *p* values index the closeness of the null hypothesis to being true, the value of the hypothesis, the degree to which the path coefficient can be trusted, or anything else of value for discerning what we should or should not deem causal. It is the height of folly that researchers use NHST to assign causation to arrows in SEM diagrams.

Using SEM to Assess Degrees of Causation

There doubtless are those who would agree that it is silly to say that causation is there or not there based on NHST, but would argue for a different use of SEM, which is to consider the extent of the causation. This argument would emphasize starting from an assumption that arrows in the SEM diagram are causal but obtaining path coefficients tells researchers how much causation there is. Path coefficients with larger absolute magnitudes, whether in the positive or negative direction, indicate a greater degree of causation than path coefficients with smaller absolute magnitudes.

This argument is certainly correct if we define the path coefficient as indicating the degree of causation; the definition renders the argument inevitable. But consider alternative definitions. For example, for interventionists, causation is demonstrated by the effect occurring when the intervention is present and not when the intervention is absent (and other criteria too). Imagine that a researcher applies SEM to correlational data and obtains a path coefficient, that we can designate as θ. From the point of view of SEM authorities, θ indicates the strength of the causal path from one variable to the next. But it is not clear what θ indicates if one switches to an interventionist perspective, which is the most obvious perspective to take if the goal is to intervene in some way to obtain a desired outcome.

Imagine a thought experiment where participants are randomly assigned to experimental and control conditions, with a subsequent dependent measure. Could we use θ from SEM to predict the size of the effect? The answer is clearly in the negative. The

experimental effect size would depend on how powerfully the independent variable is manipulated, the validity of the dependent measure, the presence or absence of causally relevant factors, and others. This is not to say that the interventionist way of thinking is necessarily correct as there are many ways to conceptualize causation. However, what the interventionist thought experiment exemplifies is that there is no sound way to translate θ into some alternative index of the strength of causation using alternative causation conceptualizations. Therefore, the use of θ to index the degree of causation is either true by definitional fiat or plain wrong.

In addition to the conceptual issue, as a matter of practical fact, there are many ways to influence path coefficients. For example, Trafimow (2021a; 2021b) showed that failure to have perfectly reliable measures, a condition that fits almost all published research in aviation, can have dramatic effects on calculated path coefficients. Depending on the pattern of reliabilities of measures, path coefficients that are there under perfect reliability can disappear under imperfect reliability. Or path coefficients that are set at zero under perfect reliability can appear, with substantial values, under imperfect reliability. Imperfect reliability can even cause path coefficients to change signs! To be sure, it is possible to use the dis-attenuation formula to attempt to correct for attenuation due to unreliability (see Gulliksen, 1987; Lord & Novick, 1968 for highly cited reviews), but this creates additional problems, and dis-attenuation has been the subject of considerable controversy (Borsboom, Mellenbergh, & van Heerden, 2004; Schmidt & Hunter, 1996; 1999; Schmidt, Le, & Oh, 2013; Zimmerman, 1975). Without going into too much detail, consider that the correction must itself be based on sample data, that may or may not accurately represent the population and is subject to random measurement error. Worse yet, if the sample reliability is low enough, correction may render correlation coefficients greater than unity. None of this is to say that not correcting is better than correcting, only that correcting is not a panacea and there are difficult issues with which to contend.

Furthermore, it is unlikely in the extreme that all variables are measured with perfect validity, or even similar validity. Just as unreliability can cause causal pathways to appear, disappear, or change signs, invalidity can create these effects too (Trafimow, 2006). The upshot is that it would take an absurd level of credulity to believe that the path coefficients in typical aviation research represent the extent of causation independent of the reliabilities and validities of the measures. Given the many factors that influence calculated path coefficients, it is unclear what they index outside recourse to definitional fiat.

Worse yet, there is the issue of whether causation is within-persons or between-persons. To approach the issue, consider that in 2019, there was a mean of 1.93 children under 18 years of age per family in the USA. However, this mean, though fine at the between-persons level, is blatantly wrong at the within-persons level because no family has 1.93 children. Now, if causation is to be expressed as a path coefficient based on a sample, that is, the typical between-persons level, it is not clear that the obtained value for θ has anything to do with causation within-persons. Nor is this a fanciful argument; Molinaar and colleagues (2004; 2008; 2015; Molinaar & Ram, 2009) argued repeatedly,

and demonstrated, that within-participants correlational analyses often fail to accord with between-participants ones. Thus, it is unclear for whom the touted causation is occurring. A caveat is that the issue of between-participants causation versus within-participants causation may influence the interpretation of true experiments too.

Configural Causation

To simplify and adopt an example from Mackie (1974), imagine that someone drops a lit match, and a fire ensues. Did the lit match cause the fire? Well, it depends. Suppose an absence of flammable material, in which case the match would not have caused the fire. Or suppose a strong wind that blows the match to where there are flammable materials, so the fire happens after all. Or suppose the presence of a person who pours water on the match before it can light the flammable materials to which the wind blew it. And so on. The point is that most causation is not as simple as *A* causes *B*, or even *A* causes *B* causes *C*, because there are many factors whose presence or absence influence whether the effect occurs in conjunction with the alleged cause. In a word, causation is *configural*. Researchers in other areas have increasingly become more sophisticated in their thinking and employed alternative sorts of analyses based on hypothesized causal configurations (Woodside, 2013a; 2013b; 2015; Woodside & Baxter, 2013). Although such analyses are not devoid of problems, they may be conceptually closer to what people really mean when they say that there is causation than is SEM.

Theory

The typical response to SEM criticisms is to chant the one-word mantra: **theory**. The mantra implies that although SEM is associated with important problems, theory comes along and makes everything okay. To avoid criticizing anyone, let us consider hypothetical cases where SEM is performed under the umbrella of a theory with previous strong or weak empirical support.

SEM with Strongly Supported Theory

Consider one of best-supported theories in the literature, the reasoned action approach (Ajzen & Fishbein, 1980; Fishbein, 1980; Fishbein & Ajzen, 1975; Fishbein & Ajzen, 2010), that has been reinforced by a variety of research paradigms, including true experiments, quasi experiments, and hundreds of SEM studies in a variety of fields, though this does not necessarily mean that the theory is true (Trafimow, 2007; 2009). For present purposes, it is not necessary to understand the whole theory, but rather a small part of it involving two pathways to behavior. There is an attitudinal pathway, whereby attitudes (evaluations of behaviors) are assumed to cause behavioral intentions which, in turn, are assumed to cause behaviors. And there is a normative pathway, whereby subjective norms (people's opinions about what most others who are important to them think they should do) are assumed to cause behavioral intentions which, in turn, are assumed to cause behaviors. Suppose that a researcher performs a SEM study and obtains an impressive path coefficient from attitudes to behavioral intentions for an aviation-

relevant behavior. How seriously should we take the finding and its associated arrow in the resulting diagram?

The arrow in the diagram indicates that this association is alleged to be causal. Absent the reasoned action theory and literature, this would be problematic because it is just as likely that behavioral intentions cause attitudes as that attitudes cause behavioral intentions, or that an outside variable or set of variables causes both attitudes and behavioral intentions. But the theory of reasoned action literature strongly supports that attitudes can cause behavioral intentions, and so the claim appears, at first glance, to be on firm ground.

However, appearances are deceiving. Consider that the theory does not insist that attitudes always cause behavioral intentions, as there is a normative pathway too. Rather, the theory only insists that *either* attitudes cause behavioral intentions *or* subjective norms do, for any single behavior of interest. And adding other reasoned action variables, such as perceived behavioral control to the mix, does not help, as the addition merely indicates that there are now three pathways to the behavioral intention. Again, the theory is agnostic as to which pathway or pathways will dominate for any single behavior of interest. Consequently, and contrary to appearances, the theory does not provide a strong prior reason to believe that attitudes cause behavioral intentions with respect to the specific behavior under investigation. And we are back to the problem that correlation need not indicate causation; it is entirely possible that the causation goes in the reverse direction or that outside variables are responsible for the obtained path coefficient. For instance, perhaps previous enjoyment of the behavior, or lack thereof, influences both attitudes and behavioral intentions. Or perhaps subjective norms cause both attitudes and behavioral intentions. Or perhaps, as Festinger and Carlsmith (1959) famously demonstrated, there was reverse causation from behavior to attitudes (possibly through behavioral intentions), thereby generating the obtained path coefficient. The data are consistent with all these possibilities, and more, thereby failing to disconfirm them in support of the touted causation.

SEM With Weakly Supported Theory

We have seen that even at its best, under the umbrella of a theory with strong prior empirical backing, SEM work is extremely problematic. The theory might specify possible causal pathways but be agnostic about which pathways work for which behaviors. In this case, even with a strong theory, there is little in the way of prior reason for believing that relationships between variables are causal in the hypothesized manner for the behavior under investigation. To paraphrase the exposition by Spirtes, Glymour, and Scheines (2000), when the data are consistent with multiple causal pathways, the support for any one of them is compromised.

But consider a more typical case, where there is a theory that had been supported by SEM studies but not by studies with stronger research paradigms. Suppose a theory specifies that variables *A*, *B*, *C*, and *D* all cause *E*. And to back this up, previous researchers have reported obtaining these relationships. Well, then, a new researcher

measures all five variables and obtains path coefficients that support the presumed relationships. How strongly should we believe the data?

It depends on what we mean by "believe the data." Given that other researchers have found that *A*, *B*, *C*, and *D* all correlate with *E*, and that the new researcher has replicated, there is good reason to believe that the four relationships exist. However, believing that the relationships are there is not the same thing as believing that *A*, *B*, *C*, and *D* cause *E*. As usual, causation could be in the other direction or by outside variables. Invoking the theory is unconvincing here. Because support for the theory is based on previous SEM empirical work, using the theory to justify the present SEM work is tantamount to saying that previous SEM work justifies future SEM work. Or more briefly, SEM justifies SEM.

Thus, we have seen that even in best-case research, where there is a strong theory with empirical support from multiple research paradigms, SEM provides only a weak form of evidence for causation. And in typical research, where there is no strong prior theory but only a litany of supporting citations to previous SEM work, new SEM work provides yet weaker evidence for causation.

Finally, there are cases where the set of arrows and constructs in a SEM diagram *is* the theory. But it is blatantly vicious to argue that we should believe the SEM diagram because of the theory when the SEM diagram is the theory!

SEM with Just the Theory

Imagine a theory that has not yet been tested. A researcher measures all the relevant variables and performs SEM on the correlational data. Without the theory, even SEM authorities would admit that conclusions about the extent of causation are unjustified, but many SEM aficionados would claim that the theory itself somehow justifies that which otherwise is blatantly unjustifiable. From this perspective, the foregoing subsections concerning strong or weak evidence for the theory could be considered to miss the point. Rather, because the theory itself specifies that which should be causal or not, performing SEM on correlational data, and drawing causal conclusions is justified.

Before addressing this issue, a quick reminder is in order. Most theories in the aviation literature are not as definite as the previous paragraph pretends. As we saw in the discussion about the theory of reasoned action, most theories specify paths that could be causal, but do not have to be. Therefore, it is useful to recall that it is one thing for a theory to demand a particular causal path, and quite another thing for a theory to permit a particular causal path. In the latter case, which is the most typical one, the obtained path coefficient, whatever the value happens to be, fails to strongly test the theory.

But let us take the highly atypical case where a theory demands an impressive path coefficient, and the path coefficient appears. There could be a temptation to say that the data strongly support the theory. But consider alternative explanations. Not only does

it remain plausible that causation is in the reverse direction or due to outside variables, but there might be competing theories that also would make the prediction. In the context of a true experiment, a researcher could handle the problem by cleverly designing the experiment so that competing theories could be forced to make opposing predictions. In principle, such clever design could be applied to correlational studies too, but I know of no aviation papers where this has been done. One possible reason for the lack is that experimental paradigms may be more susceptible to cleverness of that sort than are correlational paradigms. A second possible reason may be that the mere fact that one is performing a correlational study may prime researchers to content themselves with measuring the variables specified by the theory, whereby the fact of performing an experimental study may prime researchers into cleverness about choices of what to manipulate and how to do it, to reduce the plausibility of competing possibilities.

The foregoing skepticism should not be taken as being anti-theory. On the contrary, theory is crucial to science. However, the mere fact that SEM applied to correlational data is in the context of a theory fails to negate the ubiquitous problems that plague correlational data. Moreover, it is worth remembering that there is no guarantee that the theory is true in the first place. If the theory is false, it hardly provides a strong case for treating an obtained path coefficient as causal. And if one is not sure about the truth status of the theory, to argue that (a) the theory supports the SEM diagram and (b) the SEM diagram supports the theory is unconvincing.

External Validity

The topic of external validity has received much focus in a variety of literatures (e.g., Calder, Phillips, & Tybout, 1981, 1982, 1983; Calder & Tybout, 1999; Epstein, 1979, 1980; Lin, Werner, & Inzlicht, 2021; Lynch, 1982, 1983, 1999; Manzi, 2012; Mook, 1983; Pearl & Mackenzie, 2018; Sears, 1986; Wintre, North, & Sugar, 2001). Most researchers consider internal validity to concern the extent to which the effect can be attributed to the putative cause whereas they consider external validity to concern generalization to different cultures, contexts, operationalizations, or other deviations from the original study paradigm. The foregoing discussion about causation, including that theory fails to be nearly as helpful as is popularly believed, can be considered to indicate that SEM internal validity is unimpressive. But does SEM external validity fare better?

We already have seen that path coefficients can be influenced by a host of factors, including reliability issues, validity issues, and so on. There were other concerns too, such as ready availability of alternative explanations, the lack of clarity about the meaning of causation in the first place, and the problem that there is no reason to expect within-participants analyses to be in concert with the between-participants analyses featured in practically all aviation SEM studies. And, on top of that, there is the problem that correlation need not imply causation, even when there is a theory. From the traditional perspective of a tradeoff between internal and external validity (Campbell et al., 1963; Cartwright, 2007; Lin et al., 2021), the internal validity problems might suggest that external validity should be impressive to balance them out. But this is not so.

Considering again the host of factors that can influence the path coefficients that aviation researchers obtain, how likely is it that similar path coefficients would be obtained upon replication attempts in different cultures, contexts, operationalizations, and so on? If different measures are used, these likely would have different reliabilities or validities, thereby resulting in different path coefficients. Moreover, there is no reason to expect path coefficients obtained with one population of pilots, trainers, or airline companies, even ignoring reliability and validity issues, to be obtained with different ones.

Nor does theory come to the rescue. It could be that a researcher uses SEM to obtain a finding consistent with a theory, but that a finding is consistent with a theory is not the same as having the theory demand that finding. The example of finding of a path coefficient between attitudes and behavioral intentions is a case in point. Although the finding is consistent with the theory, it is not demanded by the theory. And this lack of demand compromises that the theory provides a sound reason to assume attitudes cause behavioral intentions with respect to the single behavior at hand. Nor does the theory provide much reason to assume that attitudes cause behavioral intentions with respect to other behaviors, cultures, operationalizations, and so on. The best that could be said is that the theory provides reasons for suspecting that attitudes might cause behavioral intentions, for various behaviors, but that is far from providing a sound reason for concluding that this must be so for the behaviors of current interest, or for generalizing across cultures, contexts, and so on.

Then, too, there is the issue of that which we wish to generalize. One answer might be that we wish to generalize the theory. But this potential answer is problematic. Firstly, if the theory is being used to justify the model that comes out of SEM, it is circular to then use the model to justify the theory. Secondly, the model is not capable of increasing the generalizability of the theory because generalization is an issue that goes well beyond statistics. Rather, demonstrating generalizability requires performing studies with different cultures, contexts, operationalizations, and so on.

Or the idea might be to generalize the models that come out of SEM. But this does not work either unless one can show that similar path coefficients are obtained with different cultures, contexts, operationalizations, and so on. It is possible to misinterpret this statement by considering generalizability successful if a statistically significant path coefficient is obtained in studies performed in different cultures. However, this is not so. For example, suppose that the path coefficient equals 0.1 (n = 1,000; $p < 0.05$) in Culture A whereas it equals 0.90 (n = 1,000; $p < 0.05$) in Culture B. Although there is statistical significance in both cultures, the coefficients differ wildly. We have already seen that significance testing applied to SEM is contraindicated, and the present example reinforces that point. Looking at the actual coefficients, instead of attending only to p-values, renders obvious that, under SEM definitional fiat, there is not much causation going on in Culture A whereas there is much causation going on in Culture B. Although most researchers would consider the example to be one supporting generalizability, due to the successful significance tests, attention to the coefficients themselves demonstrates a generalizability failure. Worse yet, if we abandon definitional fiat, then it is unclear

whether zero, one, or two of the coefficients has anything to do with causation whatsoever.

Finally, a ubiquitous generalizability issue regards sampling precision. In general, the larger the sample size, the better the sample statistics estimate the corresponding population parameters (Trafimow, 2019b). However, as McQuitty (2004; 2018) famously asserted, if the sample size is too large, then tenable models created by SEM are likely to be rejected by significance testing. Hence, McQuitty suggested that SEM researchers restrict themselves to moderate sample sizes. Although this advice is good from the point of view of not rejecting models, it necessarily reduces how well sample path coefficients estimate population ones (Trafimow et al., 2021); in general, researchers tend to use sample sizes that are insufficient to provide precise estimates of corresponding population parameters (Trafimow & Myüz, 2019; Trafimow, Hyman, & Kostyk, 2020; Trafimow et al., 2021). To the extent that researchers comply with standard SEM advice pertaining to sample sizes, such compliance serves to decrease generalizability.

In summary, external validity is always a difficult issue, regardless of research paradigms. Applying SEM to correlational data does not increase external validity unless the study is performed in different contexts or cultures, with different operationalizations of the constructs, and so on. Nor does obtaining statistically significant findings across contexts, cultures, operationalizations, and so on necessarily support generalization. It is necessary to look at the actual coefficients and determine how similar or different they are. There is never a shortcut to external validity, and going the SEM route does not provide one either.

Items I Am Not Saying

To prevent misinterpretation, it is important to be clear about what I am not saying. To recapitulate, I am saying the following about SEM applied to correlational data to arrive at causal conclusions: the SEM definition of causation is not useful; SEM causation does not translate well to other causation conceptions; abandoning causation by definitional fiat renders SEM an extremely poor grade of evidence for causation, even when there is a theory, and no matter whether the theory has received strong, weak, or no support previously; SEM fails to provide the extent of the causation, even under the generous assumption that causation is there; and SEM is no more externally valid than other research paradigms, so an appeal to external validity fails to save SEM from the aforementioned limitations. However, I am not saying the following.

1. *Confusing SEM as a statistical technique with SEM as a research paradigm*. The present claim is not that there is anything wrong with SEM as a statistical method. It is possible to conceptualize many standard statistical procedures as being subsumed by SEM. For example, one can use SEM, as a statistical technique, to analyze data obtained from true experiments too, not just correlational designs. Rather, the present complaint is with SEM as a research philosophy that claims to render causation from correlational data.

2. *Denigrating Theory*. The foregoing argument that theory fails to justify SEM is not meant to denigrate the value of theory in science. Theory is very important for a variety of reasons pertaining to explanation, prediction, control, and others. However, that theory is important, in general, is different from saying that theory justifies SEM. It is possible for theory to be important even though it does not justify SEM as it is used in aviation research, and that is the present position.

3. *Reduction to Two Choices*. Under the notion that definitive causation, in an absolute sense, is impossible to demonstrate, it might be tempting to construe the present argument in the context of a misleading dichotomy. That is, one never draws causal conclusions because the study is never definitive, or one is not thusly bound, in which case it is fine to draw causal conclusions from nondefinitive analyses. Or to put it in the form of a pointed question: If one refuses to draw causal conclusions because of the SEM limitations, should not one avoid drawing causal conclusions from experiments too, that also have limitations? But such dichotomous thinking ignores that there are degrees of quality of evidence. In a gold-standard experiment, the evidence for causation is much stronger than in SEM paradigms in aviation. Therefore, although even a gold-standard experiment is not definitive, due to the impossibility of only manipulating one construct, it might not be unreasonable to draw causal conclusions. An effect in a gold-standard experiment strongly implies that the effect on the dependent variable is due to the manipulation, though there can be arguments about precisely what construct was manipulated. For example, in the famous Festinger and Carlsmith (1959) study, the authors manipulated the amount of money paid to participants to lie about the interestingness of the experiment and interpreted this as a cognitive dissonance manipulation. In contrast, Bem (1967) interpreted it as a self-perception manipulation. Although it may not be clear that either party's interpretation is correct, there is consensus, at least, that the manipulation causes the effect. In contrast, for SEM research, it is unreasonable to draw causal conclusions of any sort, absent strong prior justification for them. And if there is strong prior justification for the causal conclusions one wishes to draw, then the SEM research cannot be transformative in the sense of instigating belief change.

4. *Claiming that Aviation Research is Uniquely Bad*. My focus on aviation research is not because I believe that aviation research is uniquely bad; rather, SEM applied to correlational data to draw causal conclusions is bad wherever it is used. It is just as bad in management, marketing, social psychology, and others, as it is in aviation. However, as the present journal is an aviation journal, there is an opportunity to use the aviation field to make important points about SEM, which is the larger goal. My hope is that aviation researchers benefit but that researchers in other fields benefit too.

5. *Being Mistakenly Unbalanced*. It is true that the present presentation focuses on the negative, but for good reason. Specifically, the sociological fact of the matter is that SEM is rapidly gaining in popularity thereby implying that people know about the benefits of feasibility and ease of use. The problem is that, as described throughout, SEM provides evidence of extremely poor quality with respect to causation, and few researchers understand just how poor the quality of evidence really is. A potential counter

is that SEM has been used successfully in other fields, so it ought to be possible to use it successfully in aviation too. However, the basic premise that SEM has been used successfully in other fields is wrong, depending on how one defines *success*. An extremely close and critical look at the zillions of publications in areas where SEM has allegedly been used successfully suggests that the truth of the matter is that the typical effect of SEM is to induce researchers to come to unjustified conclusions. A perusal of such papers suggests a potential argument that SEM results in no benefit above and beyond the correlation matrix or simple regression analyses. In fact, to the extent that SEM diagrams promote unjustified causal conclusions, it is easy to assert that including these diagrams causes harm relative to not including them. Of course, if one defines success as when the researcher obtains grants and publications, then SEM is extremely successful, but that is not the definition to which researchers ought to subscribe.

6. *Correlational Methods More Generally*. One previous reader felt that the foregoing exposition implies that other correlational methods are better than SEM. But of course, this is not so. Other correlational methods are subject to the foregoing criticisms too, but with an important exception. To see this, suppose that an aviation researcher finds that attitudes are correlated with intentions to seek pilot training. Well, then, for a zero-order correlation, there might be less of a temptation to draw a causal conclusion and fall into the traps discussed earlier. In contrast, if this relationship is part of a larger SEM diagram, that includes an arrow from attitudes to intentions, then the temptation to draw a causal conclusion is tantamount to irresistible. The arrow itself implies causation!

7. *Construct Validity Is Unimportant*. A previous reader felt that the glory of SEM is in demonstrating construct validity and that the foregoing exposition fails to recognize that, but rather presumes that construct validity is unimportant. So, let us consider construct validity. If we return to the original construct validity piece by Cronbach and Meehl (1955), they clarified that researchers obtain construct validity by demonstrating matches between theoretical and empirical relations. Over multiple studies, with each study demonstrating such correspondences, researchers in a particular domain can establish a nomological network of theoretical and empirical relations. The nomological network simultaneously supports the theory and that the measures of the constructs are valid, i.e., construct validity. Unfortunately, however, most SEM researchers seem to feel that establishing a nice factor structure with respect to indicators and latent variables in a single study is equivalent to establishing construct validity. Obviously, however, it is not unless one substantially waters down the meaning of construct validity from the traditional Cronbach and Meehl sense. A complete discussion of construct validity, including the issue of whether it is the most important type of validity, is far beyond present scope. But the present paragraph about construct validity should be sufficient to demonstrate that construct validity is not about a particular statistical technique, or even about relations between latent variables and their indicators, but about correspondences between theoretical and empirical relations. Establishing a nice factor structure involving latent variables and indicators completes only a small corner of a very large construct validity mosaic. It is unfortunate that many SEM researchers fail to understand that until they establish a nomological network of theoretical and empirical relations, they have not

demonstrated construct validity. It is very difficult to establish construct validity in a single study, whether it is a gold standard experimental study or a SEM study.

8. *Nothing We Do Is Worth Doing*. That there are many problems with the field should not be exaggerated to mean that nothing is worth doing. For the sake of intellectual honesty, I would agree that much that has been published was not worth doing, but this is not equivalent to advocating punting on third down. On the contrary, there is much that is worth doing. For one thing, there is nothing wrong with performing correlational work if accompanied by modesty about the conclusions. To reuse a previous example, if a researcher finds that attitudes and behavioral intentions to engage in pilot training are correlated, the finding may be useful even without attributing causation. The researcher might invent multiple theories to explain the finding and then test them against each other. The test might be in the form of a gold standard experiment, but it might be in the form of a correlational study too. Suppose that one explanation is that attitude causes behavioral intention whereas another explanation is that subjective norm causes both. To go the experimental route, one might manipulate attitudes and get an effect on behavioral intentions in Experiment 1 and manipulate subjective norms in Experiment 2 with no effects on attitudes or behavioral intentions. This would constitute fairly strong evidence in favor of the former over the latter explanation. But a correlational paradigm could be used too. Imagine that the researcher measures attitudes, subjective norms, and behavioral intentions and obtains a strong correlation between attitudes and behavioral intentions but not between subjective norms and either attitudes or behavioral intentions. That pattern of correlations would provide a fairly strong case that one explanation is better than the other. In turn, if one is designing an advertisement to induce people to take pilot training, the additional research, whether correlational or experimental, might provide a sufficient reason to focus on attitude change as opposed to focusing on subjective norm change. Thus, there is much that is worth doing, but it requires the difficult cognitive work of generating and testing alternative explanations for correlational findings. An unfortunate side effect of the increasing dominance of SEM is that it provides a too-good-to-be-true solution that discourages aviation researchers from generating and testing alternative possibilities.

Conclusion

Because true experiments are difficult to perform and are even more difficult to perform when one is interested in mediation, it is understandable that aviation researchers desire a more feasible research paradigm. As we saw in the introduction, SEM studies in aviation tend to be based on correlational designs where the researcher simply measures the variables of concern, with the SEM program doing the heavy lifting of providing path coefficients, significance tests, model fitness statistics, and so on. There can be little doubt that the SEM way to perform research is the easy way, relative to true experiments. But easiness is not the issue of present concern. Rather, the issue of present concern is whether SEM paradigms in aviation provide strong evidence either for causation or, once causation is assumed, the extent of the causation. Unfortunately, we have seen that neither is so. SEM analyses provide a very weak form of evidence for causation, which is why SEM authorities repeatedly state the importance of having strong prior evidence for

causation. However, we also have seen that the path coefficients are uninformative about the degree of causation because SEM-defined causation is not commensurable with other causation conceptions, such as the interventionist conception that provides the foundation for true experimental research.

Nor does it help that even when aviation researchers are cautious about making causal claims based on correlational data, they nevertheless bring causation in through the back door when making recommendations for pilot training, interventions, or policy. Without a belief in causation, there would be little reason to believe that such recommendations would work. That there is inevitably some sort of recommendation in the discussion section shows that researchers are drawing causal conclusions even when being cautious in the introduction. In addition, that SEM diagrams have arrows further indicates that causation is being touted, no matter how careful the language. This is too bad. It is understandable that when experiments are not feasible, researchers would resort to correlational research paradigms. But they should not deceive themselves about the quality of the evidence: it is extremely weak. In those case where it is possible to perform true experiments, but difficult, researchers should make the effort and perform the true experiments. Recognizing that researchers need to publish for the sake of their careers, they are unlikely to follow this recommendation. But at least they could avoid deceiving themselves, and others, by imputing causation when there is so little reason to do so. Moreover, instead of making unjustified recommendations for intervention, policy changes, and so on, researchers could expend more cognitive effort generating competing explanations that could then be tested against each other in future research, whether that research is experimental, correlational, or follows some other paradigm.

References

Ajzen, I. (1988). *Attitudes, personality, and behavior*. Dorsey.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Prentice-Hall.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychology Review*, *74*(3), 183-200. https://doi.org/10.1037/h0024835

Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg & S. Cohen (Eds). *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (2nd Ed, pp. 235–254). Aldine de Gruyter.

Bollen, K. A., & Pearl, J. (2013). Eight myths about causality and structural equation models, in S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301-328). Netherlands: Springer.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bradley, M. T., & Brand, A. (2016). Significance testing needs a taxonomy: Or how the Fisher, Neyman-Pearson controversy resulted in the inferential tail wagging the measurement dog. *Psychological Reports*, *119*(2), 487-504. doi: 10.1177/0033294116662659

Byrne, B. M. (2013). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*(2), 197-207. https://doi.org/10.1086/208856

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1982). The concept of external validity. *Journal of Consumer Research, 9*(3), 240-244. https://doi.org/10.1086/208920

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1983). Beyond external validity. *Journal of Consumer Research*, *10*(1), 112-114. https://doi.org/10.1086/208950

Calder, B. J., & Tybout, A. M. (1999). A vision of theory, research, and the future of business schools. *Journal of the Academy of Marketing Science, 27*(3), 359-366. https://doi.org/10.1177/0092070399273006

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research.* Houghton, Mifflin and Company.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Davis F. D., Bagozzi, R.P., Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982–1003. doi: 10. 1287/ mnsc.35.8. 982

Epstein, S. (1979). The stability of behavior I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*(7), 1097-1126.

https://doi.org/10.1037/0022-3514.37.7.1097
Epstein, S. (1980). The stability of behavior II. Implications for psychological research. *American Psychologist*, *35*(9), 790-806. https://doi.org/10.1037/0003-066X.35.9.790
Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*(2), 203–10. doi: 10.1037/h0041593
Fishbein, M. (1963). An investigation of the relationships between beliefs about an object and the attitude toward that object. *Human Relations*, *16*, 233-239.
Fishbein, M. (1967). Attitude and the prediction of behavior. In M. Fishbein (Ed.), *Readings in attitude theory and measurement* (pp.477-492). Wiley.
Fishbein, M. (1980). Theory of reasoned action: Some applications and implications. In H. Howe & M. Page (Eds.), *Nebraska Symposium on Motivation, 1979* (pp.65-116). University of Nebraska Press.
Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Addison-Wesley.
Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. Psychology Press (Taylor & Francis).
Fussell, S. G., & Truong, D. (2020). Using virtual reality for dynamic learning: an extended technology acceptance model. *Virtual Reality*. https://doi.org/10.1007/s10055-021-00554-x
Grice, J. W., Cohn, A., Ramsey, R. R., & Chaney, J. M. (2015). On muddled reasoning and mediation modeling, *Basic and Applied Social Psychology, 37*(4), 214-225. doi: 10.1080/01973533.2015.1049350
Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Heise, D. R. (1975). *Causal analysis*. New York: John Wiley & Sons.
Hirschauer, N., Grüner, S., Muβhoff, O., Becker, C., & Jantsch, A. (2020). Can *p*-values be meaningfully interpreted without random sampling? *Statistics Surveys*, *14*, 71-91. doi: 10.1214/20-SS129
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. https://doi.org/10.1080/10705519909540118
Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199-218. https://doi.org/10.1086/376806
Jenatabadi, H. S., & Ismail, N. A. (2014). Application of structural equation modelling for estimating airline performance. *Journal of Air Transport Management*, *40*, 25-33. doi: 10.1016/j.airtraman.2014.05.005
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.
Kenny, D. (1979). *Correlation and causality*. New York: John Wiley & Sons.
Kline, R. B. (2015) The mediation myth. *Basic and Applied Social Psychology*, *37*(4), 202-213.

doi: 10.1080/01973533.2015.1049349
Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, *16*(4), 854-863. doi: 10.1177/1745691620974773
Lynch, J. G., Jr. (1982). On the external validity of experiments in consumer research. *Journal of Consumer Research*, *9*(3), 225-239. https://doi.org/10.1086/208919
Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, *16*(4), 854-863. doi: 10.1177/1745691620974773
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Lynch, J. G., Jr. (1983). The role of external validity in theoretical research. *Journal of Consumer Research*, *10*(1), 109-111. https://doi.org/10.1086/208949
Lynch, J. G., Jr. (1999). Theory and external validity. *Journal of the Academy of Marketing Science*, *27*(3), 367-376. https://doi.org/10.1177/0092070399273007
MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323-326. https://doi.org/10.1177/0092070303031003011
Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Oxford University Press.
Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society*. New York, N. Y.: Basic Books.
McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, *57*(2), 175–183. https://doi.org/10.1016/S0148-2963(01)00301-0.
McQuitty, S. (2018). Reflections on "Statistical power and structural equation models in business research". *Journal of Global Scholars of Marketing Science*, *28*(3), 272–277. https://doi.org/10.1080/21639159.2018.1434806.
Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*, 201-218.
Molenaar, P. C. M. (2008). Consequences of the ergodic theorems for classical test theory, factor analysis, and the analysis of developmental processes. In S. M. Hofer & D. F. Alwin (Eds.), *Handbook of cognitive aging* (pp. 90–104). Thousand Oaks, CA: Sage
Molenaar, P. C. M. (2015). On the relation between person-oriented and subject-specific approaches. *Journal of Person-Oriented Research*, *1*, 34-41.
Molenaar, P. C. M., & Ram, N. (2009). Advances in dynamic factor analysis of psychological processes. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 255-268). New York: Springer.
Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379-387. https://doi.org/10.1037/0003-066X.38.4.379
RaviKumar, K., Ramanathan, K., & Porkodi, S. (2012). Measuring commuters'

expectations on aviation service quality: A structural equation model (SEM) approach. *European Journal of Scientific Research, 86*(3), 402-412. http://www.europeanjournalofscientificresearch.com

Saylors, R., & Trafimow, D. (2020). Why the increasing use of complex causal models is a problem: On the danger sophisticated theoretical narratives pose to truth. *Organizational Research Methods*. https://doi.org/10.1177/1094428119893452

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199-223. https://doi.org/10.1037/1082-989X.1.2.199

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183-198. https://doi.org/10.1016/S0160-2896(99)00024-0

Schmidt, F. L., Le, H., & Oh, I-S. (2013). Are true scores and construct scores the same? A critical examination of their substitutability and the implications for research results. *International Journal of Selection and Assessment, 21*(4), 339-354. https://doi.org/10.1111/ijsa.12044

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515-530. https://doi.org/10.1037/0022-3514.51.3.515

Singh, V., Vaibhav, S., & Sharma, S. Kr. (2021). Using structural equation modelling to assess the sustainable competitive advantages provided by the low-cost carrier model: The case of Indian airlines. *Journal of Indian Business Research, 13*(1), 43-77. doi: 10.1108/JIBR-12-2017-0260.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, Massachusetts: The MIT Press.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology, 37*(4), 235-246. doi: 10.1080/01973533.2015.1062380

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology, 37*(4), 226-234. doi: 10.1080/01973533.2015.1049351

Trafimow, D. (2006). Multiplicative invalidity and its application to complex correlational models. *Genetic, Social, and General Psychology Monographs, 132,* 215-239. doi: 10.3200/MONO.132.3.215-240

Trafimow, D. (2007). Distinctions pertaining to Fishbein and Ajzen's theory of reasoned action. In I. Ajzen and D. Albarracin (Eds.), *Prediction and Change of Health Behavior: Applying the Reasoned Action Approach* (pp. 23-42). Mahwah, NJ: Erlbaum Associates, Inc.

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology, 19*(4), 501-518. doi: 10.1177/0959354309336319

Trafimow, D. (2015). Introduction to the special issue on mediation analyses: What if planetary scientists used mediation analysis to infer causation? *Basic and Applied Social Psychology, 37*(4), 197-201. doi: 10.1080/01973533.2015.1064290

Trafimow, D. (2017). The probability of simple versus complex causal models in causal analyses. *Behavior Research Methods, 49*(2), 739-746. https://doi.org/10.3758/s13428-016-0731-3

Trafimow, D. (2019). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571-583. doi: 10.1080/13645579.2019.1610592

Trafimow, D. (2019b). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics*, *7*(2), 1-14. https://www.mdpi.com/2225-1146/7/2/26

Trafimow, D. (2021a). Revisiting old-fashioned reliability and validity concerns. *Acta Scientific Neurology*, *4*(8), 81-87. https://www.actascientific.com/ASNE/pdf/ASNE-04-0409.pdf

Trafimow, D. (2021b). The underappreciated effects of unreliability on multiple regression and mediation. *Applied Finance and Accounting*, *7*(2), 14-30. doi:10.11114/afa.v7i2.5292

Trafimow, D., Bromgard, I.K., Finlay, K. A., Ketelaar, T. (2005). The role of affect in determining the attributional weight of immoral behaviors. *Personality and Social Psychology Bulletin*, *31*(7), 935-948. doi: 10.1177/0146167204272179

Trafimow, D. T., Hyman, M. R., & Kostyk, A. (2020). The (im)precision of scholarly consumer behavior research. *Journal of Business Research*, *114*, 93-101. doi: 10.1016/ j.jbusres.2020.04.008

Trafimow, D., Hyman, M. R., Kostyk, A., Wang, C., & Wang, T. (2021). The harmful effect of null hypothesis significance testing on marketing research: An example. *Journal of Business Research*, *125*, 39-44. https://doi.org/10.1016/j.jbusres.2020.11.069

Trafimow, D., & Myüz, H. A. (2019). The sampling precision of research in five major areas of psychology. *Behavior Research Methods*, *51*(5), 2039–2058. https://doi.org/ 10.3758/s13428-018-1173-x.

Winer, R. S. (1999). Experimentation in the 21st century: The importance of external validity. *Journal of the Academy of Marketing Science*, *27*(3), 349-358. https://doi.org/10.1177/0092070399273005

Winter, S. R., Crouse, S. R., Rice, S. (2021). *Technology in Society*, *65*, 101576. https://doi.org/10.1016/j.techsoc.2021.101576 R

Winter, S. R., Keebler, J. R., Lamb, T. L., Simonson, R., Thomas, R., & Rice, S. (2021). The influence of personality, safety attitudes, and risk perception of pilots: A modeling and mediation perspective. *International Journal of Aviation, Aeronautics, and Aerospace*, *8*(2). https://doi.org/10.15394/ ijaaa.2021.1594

Wintre, M. G., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology*, *42*(3), 216-225. https://doi.org/10.1037/h0086893

Woodside, A. G. (2013a). Proposing a new logic for data analysis in marketing and consumer behavior: Case study research of large-N survey data for estimating algorithms that accurately profile X (extremely high-use) consumers. *Journal of Global Scholars of Marketing Science*, *22*(4), 277–89. doi: 10.1080/21639159.2012.717369

Woodside, A. G. (2013b). Moving beyond multiple regression analysis to algorithms: Calling for a paradigm shift from symmetric to asymmetric thinking in data analysis and crafting theory. *Journal of Business Research*, *66*, 463–72. doi: 10.1016/j.jbusres.2012.12.021

Woodside, A. G. (2015). Constructing business-to-business marketing models that overcome the limitations in variable-based and case-based research paradigms. *Journal of Business-to-Business Marketing 22*(1–2), 95–110. doi: 10.1080/1051712X.2015.1021589

Woodside, A. G., and R. Baxter. (2013). Achieving accuracy, generalization-to-contexts, and complexity in theories of business-to-business processes. *Industrial Marketing Management 42*(3), 382–93. doi: 10.1016/j.indmarman.2013.02.004

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*(3), 161-215. https://nam10.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdrive.google.com%2Ffile%2Fd%2F1Iq3vEFZna4XXzjs-JIW9mnBu7dT-muBC%2Fview%3Fusp%3Dsharing&data=04%7C01%7Cmhyman%40nmsu.edu%7C1598ba1dc3434345274408d950a89691%7Ca3ec87a89fb84158ba8ff11bace1ebaa%7C1%7C0%7C637629507291121340%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=PeF2EsPJO8pSrGsmqmrwxnm5BQ5anHbjZYtL0sc7pTA%3D&reserved=0

Zimmerman, D. W. (1975). Probability spaces, Hilbert spaces, and the axioms of test theory. *Psychometrika, 40,* 395-412. https://doi.org/10.1007/BF02291765