Controlling (for) first-order task performance in studies on metacognition: cons & cons

Borysław Paulewicz & Marta Siedlecka, C-Lab, Jagiellonian University, Kraków

Tel Aviv, 2021

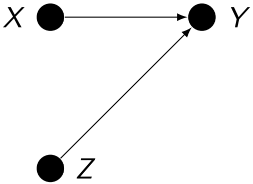
Imagine that you see an effect in some measure that you hope captures something metacognitive but you also see an effect in type 1 or first-order task performance.

Many people seem to believe that this is automatically a problem in the sense that the former effect may be in some part **due to** the latter.

If this is a problem it is a problem of *alternative causal explanations* of the observed *statistical effects*, i.e., a **confounding** problem.

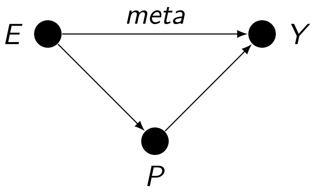
Confounding is any reason that $p(y|x) \neq p(y|do(x))$.

Just because something influences y does not mean that it is a confound of an effect in y, for example:



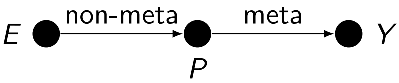
Here, Z is not a confound (with respect to estimating $p(Y|do(X))$ by regressing Y on X), Z is just "noise". Controlling for Z will not introduce bias – it may increase power, but it is not essential.

If you are faced with this kind of situation:



where E is some experimental manipulation, Y is some measure you like, and P is either first-order / type 1 task performance or first-order / type 1 sensitivity then if this (weird) model is linear estimating the direct effect of E on Y is the same as estimating $p(Y|E, P)$. If the model is nonlinear you cannot escape counterfactuals (see Pearl on mediation analysis) but you can still estimate the direct effect.

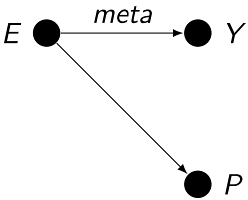
If you are faced with this kind of (two-stage) situation:



then estimating $p(Y|E,P)$ will show no effect of E on Y because of complete mediation. You could divide $R^2_{E,Y}$ by $R^2_{E,P}$ or simply estimate $p(Y|P)$ to obtain the causal effect of P on Y.

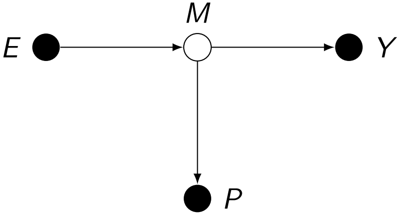
There is no confounding here.

If you are faced with this kind of situation:



then even though every two variables are correlated there is no confounding and you don't have to do anything. In fact, it would be best if you did not control for P at all in this situation.

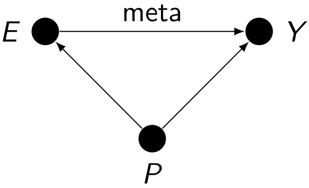
If you are faced with this kind of situation (another two-stage model):



then estimating $p(Y|E, P)$ to capture any arrow or path will be a serious **error** (that I will illustrate with a simple simulation in R if asked).

This shows that in general, you should *not* try to control for arbitrary correlates of the variable of interest.

What you cannot be faced with because it is impossible (E is randomly assigned) is this model:



Note how previous models where P is not a mediator imply that P does not have to be or must not be controlled for.

Now, task *performance* is a relation between the responses and the **objective** stimuli (e.g., observed or true PC). Typically, *performance as such* can *influence* some property of a person only when there is feedback (which is caused by performance by definition).

That is why the models with $P =$ performance as a mediator are quite weird.

We have established that controlling for *first-order / type 1 task-performance* may be unnecessary, harmful, or tricky, but what about controlling for type 1 *sensitivity*?

Type 1 tasks, type 2 tasks, and the unbearable frivolity of the meta-d' model

Some of you perhaps think "metacognitive" when you hear "type 2" and "non-metacognitive" or "underlying" or "first-order" when you hear "type 1".

If that is the case it is probably the fault of some authors whom I will quote later.

By some authors I mostly mean Fleming, Lau, and Maniscalco (in alphabetical order).

According to Clarke, Birdsall, and Tanner (1959) who introduced the terms ”type 1” and ”type 2”, or according to Galvin, Podd, Drga, and Whitmore (2003) who strongly influenced the authors of the highly influential meta-d’ model:

A type 1 task is any task in which the participant is only required to respond to some property of the stimulus.

A type 2 task is any task in which the participant is required to assess the correctness of her type 1 response.

The type 1 vs type 2 distinction is about the **task**, *not* about the *required kind of cognitive processing*. There is no such thing as a type 1 or a type 2 *process*.

A type 2 response may be influenced by a metacognitive process but it does not have to be because typically **the response**, once emitted, **is an external event**, just like the external stimulus that lead to the response.

Note however, that Galvin et. al. consider a (confidence) rating type 2 task which is a *special kind* of type 2 task that does seem to require metacognition.

A type 1 response may also be partially influenced by some metacognitive process. In fact, **the main function of metacognition is to regulate task performance.**

Galvin et. al., showed that:

if we assume an SDT model of decision making

then Type 2 sensitivity **is a function of** (not "is influenced by") type 1 sensitivity and type 1 criterion.

There is only **one** kind of evidence sample (or internal stimulus-related information) in this model and so **there is nothing metacognitive about this model**.

(Maniscalco and Lau (2014) provide an excellent analysis of this model)

Before we go any further let me say that I think

Maniscalco & Lau (2014) *Signal detection theory analysis of type 1 and type 2 data: meta-d', response-specific meta-d', and the unequal variance SDT mode*.

is a really great chapter. There is also some good stuff in

Maniscalco & Lau, (2012) *A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings.*

However

A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings (Maniscalco & Lau, 2012)

"We argue that currently available methods [of measuring metacognitive sensitivity] are inadequate because they are **influenced by** factors such as response bias and type 1 **sensitivity** [...] Extending the [...] approach of Galvin, Podd, Drga, and Whitmore (2003), we propose a method of measuring type 2 sensitivity that is **free from these confounds**".

"[...] while these [other than meta-d'] approaches are simple conceptually and computationally, they do not model type 2 sensitivity and type 2 response bias **as separate processes** and thus risk **confounding** them."

This is fine: "Clarke et al. (1959) and more recently Galvin et al. (2003) discussed how distributions of evidence for correct and incorrect stimulus judgments could be **derived from** the type 1 SDT model. An important lesson from this work is that type 1 sensitivity (d0) and response bias (c1) **influence** the [observed] area under the type 2 ROC curve (Fig. 1B)."

This is *not* fine: "This entails that two metacognitively optimal observers could differ on type 2 performance **due only to** differences in **type 1 performance**."

This is *not* fine: "Suppose observer A has d = 1, c1 = 0 and observer B has d = 2, c1 = 0, but that both observers make optimal use of the **type 1 information** [=type 2 information in the Galvin et. al. model] available to them when performing the type 2 task. B will have greater area under her type 2 ROC curve than A, and in general her confidence ratings will be more predictive of accuracy. In this sense, B has greater "absolute" type 2 sensitivity than A. But by hypothesis, the difference in their **metacognitive performance** derives entirely from informational differences at the type 1 level, and so in a sense it is misleading to conclude that the **metacognitive mechanisms** [huh?] of B are operating at a higher level of efficiency or sensitivity than those of A. The difference in their absolute type 2 sensitivity reflects the difference in **the quality of type 1 information they are metacognitively evaluating** [this is a big leap], rather than in the quality of the evaluation itself."

To summarize:

- Performance was equated with efficiency or sensitivity (this may be an easy fix).
- A functional relation was equated with a causal one.
- A type of task (1 vs 2) was equated with type of information or type of cognitive process (purely non-metacognitive / first-order / underlying vs. metacognitive)

No wonder that instead of:

”Sine 2012 me (meaning the second author) and some of my colleagues have been writing about this idea that type 1 sensitivity and type 1 criterion is a confound in some studies on metacognition and we have been quite succesful in promoting a particularily speculative solution to this alleged problem”.

this is the *first sentence* in Rahnev & Fleming (2019):

”It is becoming **widely appreciated** that higher stimulus sensitivity **trivially** increases estimates of metacognitive sensitivity.”

I could go on like this for much longer but I have only 8 minutes.

Anyone who tries to follow the references provided by the main proponents of controlling for "type 1 performance" when measuring metacognition will soon discover that it is rather difficult to see how wide this appreciation is.

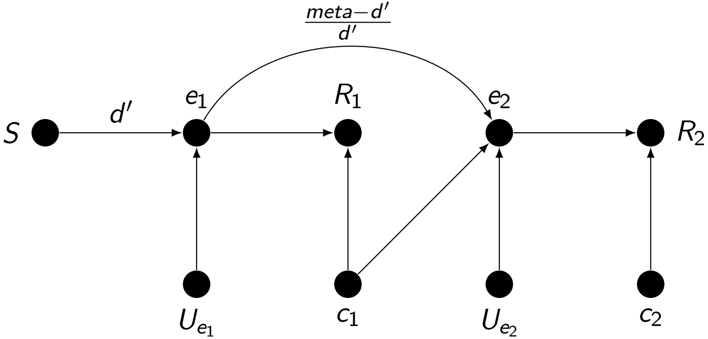
To arrive at the meta-d' model suppose that:

1. Type 1 response is *purely first-order*, i.e., it is not influenced by anything metacognitive (no metacognitive regulation is involved)
2. Type 2 response uses *only* the information on which the type 1 response was based (no metacognitive monitoring beyond retaining the information already acted upon is invloved) but it is (somehow) a metacognitive response.
3. An SDT model correctly captures the $S \rightarrow R_1$ effect.
4. Type 2 response, *conditional on type 1 response and type 1 criterion*, is distributed as if a *counterfactual* (so nonexistent by definition) *ideal observer* was doing the responding.

This way the meta-d' parameter captures (some form of) the remaining effect of S on R_2.

Exactly one of the four assumptions is not highly problematic.

Since confounding is a causal notion here is my attempt to capture the causal assumptions of the meta-d' model implied by the way the model is described and fitted by Fleming, Lau, and Maniscalco:



where $e_1 = \mu_S + U_{e_1}$, $U_{e_1} \sim N(0, 1)$, and $R_1 = 1$ *iff* $e_1 > c_1$

To summarize:

The meta-d' model is not a process model (or a "generative" model), it is not *derived from* an SDT model, nor is it firmly based on a "principled SDT framework".

It is not even remotely true that "the properties of the meta-d' model have been thoroughly explored" (Fleming, 2017).

In every simulational study (e.g., Barret, Dienes, & Seth, 2013) that showed that some *implementation of the meta-d' model fitting procedure* is perhaps fine the *validity of the main assumptions of the model was not seriously questioned*.

The meta-d' model is not useful as a model of metacognitive anything except maybe for some extremely special situations.

What about *physically* controlling type 1 task performance by staircasing or calibration?

Fortunately, this is a one-slide problem.

- Staircasing or calibration *make the differences in type 1 task performance* ***disappear from the percent-correct scores*** but, since there is no magic involved, the differences reappear in stimulus intensity.
- Staircasing or calibration may or may not change how the underlying cognitive process works but confounding has everything to do with how the data generating process works.
- If there is metacognitive regulation involved in the type 1 task then the type 1 performance should definately not be equalized.
- A staircased task is metacognitively a very weird task because of the **induced artificial relation between task difficulty and effort**.

My general advice on controlling (for) type 1 or first-order performance in studies on metacognition is *just don't do it unless you clearly understand why you have to*.

If you decide to do it, do it for the right reasons.

If you decide not to do it, don't do it for the right reasons.

The right reasons in this case are causal in nature.