

At least one of the papers that James Damore cited is bogus

James Thomas

Copyright 2017 James Thomas

News has just broken that a software developer at Google circulated a memo he'd written containing putatively scientific assertions to the effect that women are innately inferior to men in certain respects (but only "on average," so presumably we mustn't hate him too much) and recommendations to strive less for diversity in the workplace. So he was fired, and the press reacted predictably: those who call themselves liberals applauded the punishment of a heretic, and those on the right shook their heads at the intolerance of scientific "truth."

Which side is right? In the truest sense, neither is, because the problem is deeper than the irritation has let on, and its implications are more sinister. But here we focus on just one of the papers cited during the media storm, because even a nonscientist can show that it's bogus.

The paper, by Pasterski *et al.*, is available at <http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8624.2005.00843.x/abstract>, and its abstract can be summarized simply. The experimenters took a group of children whose ages ranged from three to ten, some of whom were encouraged more than others to play with sex-typical toys, and some of whom have congenital adrenal hyperplasia (CAH), a disorder that virilizes a child prenatally, *i.e.*, makes the child more biologically male than normal (on boys it has little effect, but in girls it can make the genitalia ambiguous to the point of "pseudohermaphroditism"). The idea was to see if these three factors—CAH, encouragement, and biological sex—influence the child's preference for sex-typical toys.

Before discussing the results, let's recast the situation in the standard language of the scientific method, such as we find in high school or middle school science courses. It will shed light on the study's mistakes. First, there are the three *independent variables* ("IVs"):

- *IV 1*: the sex of the child. This can take two values: male and female.
- *IV 2*: whether or not the child has CAH. This can take two values: the child either does or does not have CAH.
- *IV 3*: how much the child was encouraged to play with sex-typical toys. This can take two values: the child received either normal or extra encouragement to play sex-typically.

And for complete clarity, let's spell out the eight possible combinations of the IVs' values:

1. A boy without CAH is given normal encouragement to play with sex-typical toys.
2. A boy without CAH is given extra encouragement.
3. Boy with CAH, normal encouragement.
4. Boy with CAH, extra encouragement.
5. Girl without CAH, normal encouragement.

6. Girl without CAH, extra encouragement.
7. Girl with CAH, normal encouragement.
8. Girl with CAH, extra encouragement.

Recall that the experimenters sought to measure how the child's behavior (his or her toy preference) varied with changes in these IVs. This is the *dependent variable* ("DV"): whether or not the child's preferred toys are sex-typical. So the DV can take two values: sex-typical or sex-atypical.

In order to perform the experiment thoroughly, *the DV has to be measured for all eight possible combinations of the IVs*. Interestingly, however, the experimenters failed to do so. Instead, their abstract reports the following:

- The boys were given normal encouragement, and those with CAH showed no more likelihood of sex-typical play.
- For girls:
 - Those without CAH were given normal encouragement, and responded with a certain degree of sex-typical toy play.
 - Those with CAH were given extra encouragement (for sex-typical play, of course) and responded with *less* sex-typical play than the other girls did.

In other words, *the DV was measured only for combinations 1, 3, 5, and 8, above*. In the case of the boys (combinations 1 and 3), the experiment was *controlled*, in that IV 3 was kept *fixed* while IV 2 was varied. For this reason, the finding of no change in the DV does indeed support the hypothesis that CAH does not affect a boy's behavior, as long as he is given no extra encouragement, and assuming that nothing more is wrong with the experiment (plenty more is indeed wrong with it, but there isn't enough space here to treat those more fundamental errors).

But for the girls (combinations 5 and 8) the experimenters did *not* control the experiment. In fact, they botched it in a way that suggests a political agenda. Consider, for example, a researcher with reductionist bias. He or she might botch the experiment by measuring the DV only for combinations 5 and 7, or only for combinations 6 and 8. If the DV showed that behavior varied with CAH, then the finding would support preconceived notions that behavior (gender) is determined by biochemistry alone—the “nature” side of “nature-vs.-nurture.”

Conversely, a researcher with social constructionist bias might botch it by measuring the DV only for combinations 5 and 6, or only for combinations 7 and 8. Again, if the DV showed that behavior varied with encouragement, then the finding would support preconceived notions that behavior is learned—this is the “nurture” side of the debate.

Now the correct way to perform the experiment, the way that tries to cancel the researchers' biases, is to measure the DV for all four combinations—5, 6, 7, and 8, since we're now restricting our attention to the girls—and to try to make sense of the results. If they came out, say, as follows

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
5	yes	normal	typical
6	yes	extra	typical
7	no	normal	typical
8	no	extra	typical

then one would conclude that neither CAH nor encouragement determines behavior. The next implication would be that, if behavior is indeed determined, then it must be determined by something else; other variables would need to be sought as candidate determinants.

If, on the other hand, the results came out as, say

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
5	yes	normal	atypical
6	yes	extra	atypical
7	no	normal	typical
8	no	extra	typical

then they would support the hypothesis that behavior is determined by CAH and not by encouragement. “Nature” would in this case win out over “nurture.”

Before moving to the next scenario, it’s worth considering this unusual one:

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
5	yes	normal	typical
6	yes	extra	typical
7	no	normal	atypical
8	no	extra	atypical

It would be a surprise, because although it would support reductionist predictions, it would also support the strange hypothesis that absence of CAH—*i.e.*, normal sexual development—produces atypical behavior. Such a result should immediately arouse suspicions of a mistake in data collection. Whether or not that turned out to be the case, the experiment should of course be repeated many times, to be sure that the results were not a fluke. Again, this unusual scenario is included here only to keep your mind alert to “edge cases” and to give you a larger sense of the way of thinking.

Finally, if the results came out like so:

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
5	yes	normal	atypical
6	yes	extra	typical
7	no	normal	atypical
8	no	extra	typical

then they would support the hypothesis that encouragement, and not CAH, determine behavior. “Nurture” would win out over “nature.” But it’s also important to appreciate that the same hypothesis would also be supported if the results were “inverted,” like so:

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
5	yes	normal	typical
6	yes	extra	atypical
7	no	normal	typical
8	no	extra	atypical

In this case, not only would we find support for determination by the environment alone (“nurture”), but we would call the results counterintuitive, because they would show that extra encouragement in fact *discourages* typical behavior. An explanation would therefore be in order. One possibility is that the children in the sample are rebellious, are somehow in the habit of defying their parents’ expectations. It would of course then be reasonable to repeat the experiment with a completely different sample of children. If the

same results obtained, and if they keep obtaining with other samples of children, then we might suspect that something in the culture being studied makes children systematically defiant of their parents' wishes.

Another possibility, which anticipates the deeper critique of this whole line of "research" (a critique I don't have space here to mount) is that the children are not rebelling, but are tuning in to something *else*, an independent variable other than IV 1, IV 2, or IV 3. Such a possibility would mean that, rather like the first example above, the experiment is searching for the wrong effect.

And now for what Pasterski *et al.* actually did. To be sure, they didn't botch the experiment in the naïvely reductionist way, and they didn't botch it in the naïvely constructivist way. The fact that they didn't test all possible combinations of the IVs (5 through 8, inclusive) shows that they did botch it. But they did so in a telling way, for they measured the DV only for combinations 6 and 7, like so:

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
6	yes	extra	atypical
7	no	normal	typical

The appearance in this table of all four values of IV 2 and IV 3 makes it look as though all four combinations have been tested. But of course they haven't. Moreover, the speciousness is *convenient to reductionists*, because the first row of this table (combination 6) makes it look as though the experimenters, by giving extra encouragement for typical play in CAH girls, went to the trouble of "cancelling out" any possible environmental influences, whereas they have not in fact done so. To have done so, they would have measured the value of the DV for combination 8 and shown it to us. If the results had turned out like so

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
6	yes	extra	atypical
7	no	normal	typical
8	no	extra	atypical

then the reductionists would lose, because it would show that something about the encouragement itself is causing atypicality. If, on the other hand, the results turned out like so:

combination	CAH (IV 2)	encouragement (IV 3)	behavior (DV)
6	yes	extra	atypical
7	no	normal	typical
8	no	extra	typical

then the reductionists would be supported, though of course they would need to show, by much repetition of the experiment, that the effect is reproducible—that it is, as it were, real.

Many scientists—more importantly, many *nonscientists* whose common sense hasn't succumbed to blind trust in so-called "experts"—would of course agree with the foregoing assessment of the study's errors. But the sad fact is that we live in an age in which even such people miss the much deeper problems with the study, if not with the whole field. To see what I mean, consider again the three independent variables. IV 1 is biological sex, which the study models as having two possible values, male and female. This is hardly a problem, since in the vast majority of people the 23rd chromosome pair is either XX ("female") or XY ("male").

But IV 2, which indicates whether or not the child has CAH, is not quite so simple. Consider this quote from Wikipedia:¹

¹https://en.wikipedia.org/wiki/Congenital_adrenal_hyperplasia, as of August 20, 2017.

Further variability is introduced by the degree of enzyme inefficiency produced by the specific alleles each patient has. Some alleles result in more severe degrees of enzyme inefficiency. In general, severe degrees of inefficiency produce changes in the fetus and problems in prenatal or perinatal life. Milder degrees of inefficiency are usually associated with excessive or deficient sex hormone effects in childhood or adolescence, while the mildest forms of CAH interfere with ovulation and fertility in adults.

In plainer language: there are different degrees of CAH, and the disorder manifests itself in different ways in different people. Thus the meaning of CAH varies, both qualitatively and quantitatively. And yet the study treats it as a single two-valued variable. Such a model coarsens the reality, from a multi-dimensional continuum (if that itself is not too scientific an oversimplification) to a binary choice. By contrast, note that there is no such flaw in modeling, say, the concentration of a solute, or the location of a particle with respect to some coordinate system. Real sciences, such as physics and chemistry, suffer neither ambiguity of definition nor difficulty of measurement.

But nevermind, because these problems are nothing compared to the problems with IV 3, which *should*, in a sane society, take people's breath away. Recall that it is supposed to measure the amount of encouragement given to the children. Working our way from smaller to bigger problems, let's begin by considering that word "measure." How, pray tell, can this be done? One can only imagine that our so-called experts put parents and children in an observation room filled with toys and counted the number of times they heard them say "Good girl!" But such a method is so fatuously stupid that one doesn't even know how to begin enumerating its flaws. For starters, how can such a count be reliable, given that the parents and children are also interacting *outside* the "laboratory"? Secondly, what about "feedback" that comes from persons *other* than the parents? Since these children's ages range from three to ten, they are constantly being exposed to messages from school, from other family members, from friends and neighbors, and, perhaps most overwhelmingly, from the media. What kind of idiocy is it to think that this nonstop barrage of influence can be magically waved away by the fiat of a "researcher," and replaced by a stupid count of the number of times the parent says, "Good girl"?

And speaking of age, has nobody noticed that it is in fact a *fourth* independent variable, and is being left *uncontrolled*? This lapse is even more ironic given that age is the most "ontologically stable" variable of all: unlike the others (even biological sex!) there is absolutely no ambiguity in its definition, no difficulty in telling how old a child is. (Just to be coy about it: even if the child's age were to suffer relativistic twin-paradox modifications owing to super-fast space travel, its age would still be well defined in principle and easy to measure in practice.) So why is nobody shocked by the fact that it wasn't controlled?

But all of these glaring incompetencies are *still* nothing when we consider one of the deepest, most basic, and most obvious, of facts about communication, namely, that *its meaning lies not in what is literally said*, but what is conveyed in every way *other* than a literal transcript of the words spoken. As the very same corps of so-called experts has gotten us scientifically to say, "Communication is 90% nonverbal." (So as not to exhaust ourselves, let's overlook the fatuity of assigning a percentage, let alone the fatuity of modeling "communication" as a simple quantity.) Take any slur, for example, be it racial, ethnic, or sexual. Say the word the wrong way and you could incite a riot. But say the same word under the right circumstances and you could easily seduce a person into intimacy, might even excite sexual desire. Indeed, if the minimum age of this sample weren't three years, one could easily argue that the subjects must be even *more* sensitive to "nonverbal communication"—for the simple reason that they haven't yet learned the language. No matter: we have the example of dogs. Speak to one in any language, and it will, all things being equal, respond the same way to the same *tone*, to the same *body language*. So although our brave pseudoscientists might indeed hear a parent speak *words* of encouragement to their children, how can they know that it will have the effect they expect? The very fact that a parent knows that the child is sexually abnormal obviously means that the parent is more likely to be *anxious* to make the child "gender-typical." Are we really supposed to believe that the child will not tune in to such anxiety, in the parent's voice, face,

or body? Are we really supposed to believe that these qualities of voice—indeed, this *unaccounted-for variable*—will not produce counterintuitive effects? And even if such foolery is to be countenanced, has no one noticed that the experiment is botched by the very fact that it isn’t *blind*, that the parents actually *know* which kids have CAH and which don’t?

Does the public really need to be reminded of such truisms? Does it really need a special license to call counterfeit “experts” out on their nonsense? Don’t be surprised if they counter—with a straight face, no less—that such effects cannot be real because they cannot be measured. Or, alternatively, they may acknowledge that the effect is real, thank us for pointing it out, and come back saying that they’ve increased the number of independent variables, by including in the experiment a spectral analysis of the parents’ voices, to somehow detect whether or not the praise is sincere. In vain is it pointed out that the thing that counts here is the child’s *experience* of the parent’s voice, let alone all the other “variables.” And experience, by its very nature, is not something that can be measured.

Such simple, common sense exposures of the radical idiocy of this approach shows that we are light-years away from the straightforwardness of counting the number of electrons in an atom, or measuring the distance to a planet. This is not science. It is derangement and delusion. And very unfortunately for those of us who are sane, it is sanctioned, funded, and empowered by the highest levels of government. Financial profit is the ultimate motive here, and as the sane reader can see, it is unaccountable to reason. When is the world going to become outraged by such institutionalized incompetence?

If you’re shocked by the errors in this study, then I regret to inform you that the scandal is actually worse by orders of magnitude. For starters, that the authors’ affiliations show them to belong to so-called “prestigious” institutions: Columbia, *etc.* Considering the high regard given these places by government and society, we must suppose that work at so-called “lesser” schools and hospitals is even shoddier.

Secondly, consider that the paper was actually published. Presumably this means it went through a process of peer review. That is, the authors submitted their writeup to the scrutiny of a couple (or a handful) of anonymous colleagues drawn from a pool of “behavioral scientists” stretching across the globe. That these anonymous reviewers approved the paper for publication means they are incompetent too.

Thirdly, this last conclusion is actually *not* surprising, for the habit of errors and incompetence was noticed several decades ago, by a number of scientists who went to the trouble of raising alarms about it—but were systematically ignored. They include people such as Ruth Hubbard, Richard Lewontin, Leon Kamin, Steven Rose, Stephen Jay Gould, William Byne, and no less a genius than Richard Feynman. As for the problems of peer review, none other than Michael Crichton became so alarmed by its tendency to establish “mafias” of mutually reinforcing incompetences that he devoted considerable time and energy to writing and publish critiques of it—again, only to be ignored. (Why? Think!)

The scandal doesn’t stop there. In fact, it involves you and me as well. To understand what I mean, consider that the problems with this study run much deeper than that of botched application of the scientific method. We can begin appreciating them by reflecting on the scientific method itself. In particular, you might already have noticed that its principles, which we just illustrated above, are based on nothing but common sense. They don’t require any calculus, any algebra, any abstruse terminology (unless we call the word “variable” abstruse), or any arcane mathematics. They require no special training at all. The only skill we used to evaluate the authors’ work was elementary *logic*, which might better be called simple common sense.

And yet we go about our lives *trusting* these so-called “scientists,” trusting that they know what they’re doing, that their internal processes of quality control actually work, and that the media would never report results of theirs that are wrong or stupid. Most importantly—this is in fact the most important point to take away from this essay—we go about our lives feeling that we ourselves are in no position to evaluate their work. Again, this is true of many academic fields, including real sciences that require proper training. But

the previous section shows that it is not true of so-called behavioral science. And that too is not surprising: subatomic particles are not part of our everyday lives; whereas behavior has been observed, engaged in, reflected on, and talked about, for many millenia. If it isn't stupid to talk of an "expert" in behavior, then why did it take until the twentieth century for the notion to arise and be made something other than laughable?

This question turns out to be too deep for this small essay; the reader interested in pursuing it will need to study some history, beginning perhaps with A. N. Whitehead's *Science and the Modern World* and Jacques Barzun's *Science: The Glorious Entertainment*. Those wanting to appreciate the faultiness so endemic to so-called behavioral science can look at Feynman's "Cargo Cult Science." And Jeffrey Masson, particularly in his *Against Therapy*, mounts careful attacks on the constellation of shams known as "psychotherapy."