

The New, Highly Touted Study On Hormones For Transgender Teens Doesn't Really Tell Us Much Of Anything

If you take a closer look, you'll see that the hyped-up headlines are unwarranted

JESSE SINGAL

FEB 07, 2023



166



60



Share



The NEW ENGLAND JOURNAL of MEDICINE

I'm choosing not to put a paywall between readers and my critique of this paper, because I think it's important that critical perspectives on it from non-hyperpartisan sources be aired far and wide. This is very labor-intensive work, though, and I believe I'm one of the only journalists covering the subject in this manner, so if you find what follows useful, please consider becoming a paid subscriber or giving a paid subscription as a gift.

anthonyholgate@yahoo.com

Subscribe

For those of you who already pay for my newsletter: Thank you! You're the reason I'm able to write what I write.

If you haven't read [Part 1](#) yet, you should do that before you read this.

As I noted in that post, earlier this month *The New England Journal of Medicine* published a highly anticipated study called "[Psychosocial Functioning in Transgender Youth after 2 Years of Hormones](#)." The research team has spent years following a cohort of kids who have been administered puberty blockers or hormones at four participating clinics. In this study, they reported on how the kids who went on hormones did over the two-year span following the start of that process. The participants filled out surveys every six months on issues pertaining to their mental health, gender dysphoria, and so on. According to the authors, the kids showed key improvements two years later. "Our results provide a strong scientific basis that gender-affirming care is crucial for the psychological well-being of our patients," said [Robert Garofalo](#), one of the principal investigators for the study, as well as co-director of the youth gender clinic at Lurie Children's Hospital in a Chicago, in a [press release](#) accompanying the study. A [number of media outlets](#) echoed this narrative.

But that's a questionable interpretation of the results. In my last post, I pointed out something arguably suspicious about the protocol, including a [version](#) that they submitted into which researchers hypothesized that members of this cohort would improve on eight measures, including ones that are just about as important to gender researchers as important outcomes, such as gender dysphoria, self-harm. Then, in the published *NEJM* paper, the researchers reported that the hypothesis and six of those variables were nowhere to be found in anxiety and depression — moved in a positive direction for natal males but not trans girls (natal males). The researchers reported that the hypothesis and six of those variables were nowhere to be found without explaining how they picked them (two improvements for natal males, one just for trans boys).

I won't rehash the whole post here, but in my view this missing variables issue does call the entire effort into question, simply because if many of the variables the researchers tracked *didn't* improve, or even worsened, the fact that they were able to cherry-pick five that did show some improvement might not mean anything at all. We may well be looking at nothing but statistical noise — we just can't say for sure since the researchers are obscuring so many of their results.

For this post, though, let's temporarily set aside this potentially crippling issue. Let's imagine, instead, that the authors had preregistered an interest in the five variables they did report all along, and let's proceed accordingly in our evaluation of their study. For the sake of this post, I'm also not going to quibble with, or even deeply evaluate, the specific statistical techniques the authored employed: As I'll show, even if we grant that they made the correct decisions here (which may or may not be the case) and take their findings at face value, the results are still ambiguous at best.

A tiny bit of political throat-clearing before we start: If you write critically about youth gender medicine, you will hear from a lot of people who are aghast that you could do so given the threats trans people in the United States (adults and children alike) face. And as Dave Weigel noted in a Friday article in *Semafor*, [Donald Trump just publicized a very crazy new proposal](#) to severely curtail both transgender rights for adults and access to youth gender medicine for kids and teens.

At the risk of [repeating myself](#), I am opposed to the sorts of policies Trump is proposing — both outright restrictions on youth gender medicine and his even more radical proposal to codify into federal policy a ban on even *adults* changing their legal sex. That latter part, in particular, is downright cruel and pointless, other than being red meat for the evangelical voters he is hoping to court in 2024. (Trump's stances on these issues shift, sometimes jarringly, with the political winds. Weigel writes that "he was best known in 2016 for publicly moderating the party's faith-infused stance on LGBT issues, including opposing North Carolina's 'bathroom bill' and inviting Caitlyn Jenner to use [whichever facilities she preferred](#) when on Trump properties.")

But as Weigel noted in his write-up, there's a pretty noteworthy difference between what Trump is proposing and the more substantive, mainstream debates currently

Discover more from Singal-Minded

A newsletter about science, social-justice-activism, why the sometimes fight, and how to help them get along better -- plus a good deal of other, more random stuff.

Over 44,000 subscribers

Subscribe

Continue reading >

Sign in

raging over youth medical transition. At the end of the day, given the rapidly increasing popularity of these treatments and the sometimes overconfident proclamations made about them by their advocates, the questions surrounding youth gender medicine are in urgent need of answers regardless of who the president is or what threats the LGBT community faces. *Now is not the time to discuss this* is not an argument — it's a derailing tactic. And it's one I encountered long before Republicans latched on to this issue, to be honest. If we wait until there are no longer reactionaries trying to profit off of fear of transgender people before we figure out exactly whether and to what extent youth gender medicine works — issues which remain, by all preexisting and widely agreed-upon standards of medical evidence, unresolved — we'll continue flying blind.

And that's all I'm going to say about that — I am a science writer more than a pundit, and if I pepper every paragraph with parentheticals reiterating that I have the “right” beliefs about Trumpist policies, this will quickly become unreadable. Plus, it doesn't really matter: What follows is correct or incorrect on its own merits, regardless of the beliefs of the author.

To be clear, this *New England Journal of Medicine* study is a significant improvement over *what passes for research* in the area of youth gender medicine (though that's a low bar to clear). It's excellent that researchers are closely following cohorts of kids going on blockers and hormones, and collecting rich data on their mental and physical health trajectories. It's also useful that this team preregistered its protocol. But the fact is that this particular study really does not provide substantive evidence that hormones improve the mental health of trans kids.

This post will be organized around the following main points:

- 1) The kids in this study had an alarmingly high suicide rate.
- 2) Most of the improvements the cohort experienced were small.
- 3) It's impossible to attribute the improvements observed in this study to hormones rather than other forms of treatment that took place at these clinics.
- 4) The one bigger improvement was in a variable that might not mean all that much.
- 5) The researchers don't even consider the possibility these treatments don't work — their only answer is “more hormones.”

The Kids In This Study Had An Alarmingly High Suicide Rate

While the authors had other issues with transparency in their study, they do note, right in the abstract, that two participants died by suicide. In the body, they write that “one [suicide occurred] after 6 months of follow-up and the other after 12 months of follow-

up.” So within about a year of starting hormones, two of this study’s 315 kids were dead. They also note that there were 11 instances of “suicidal ideation during study visit.”

Event	No. of Events in Sample
Any event	15
Death by suicide	2
Suicidal ideation reported during study visit	11
Severe anxiety triggered by study visit	2

Let’s set aside the ideation issue, because the researchers have not provided us with the information we need to evaluate it. “Suicidal ideation” can mean very different things, ranging from occasional, fleeting thoughts of suicide to, much more seriously, the presence of a plan and the possession of whatever tools are required to carry it out. “I do think it’s fair to say that their use of ‘suicidal ideation’ is ambiguous,” said a suicide researcher whose brain I have sometimes picked on this issue, but to whom I always offer anonymity because they are totally uninvolved in the youth gender medicine fight. The *NEJM* researchers did administer a suicidal ideation *scale* to capture this in richer detail, but as we know from the last post, they simply didn’t report that data. So there’s just no way to know whether the 11 incidents of suicidal ideation being reported during visits with the researchers is high or low or somewhere in the middle.

As for the rate of completed suicides, a common way to measure and compare suicide rates and other such outcomes is per 100,000 individuals, per year. In the US that figure is 13.9-ish, though of course it can be recalculated every year and varies significantly by subgroup. The closest we can get for an annual estimate for the *general* population within the age range of the study (12–20) is 14.2 suicides per 100,000 members of the 15–24 age band.

As Michael Biggs, a sociology professor at Oxford University and a frequent critic of youth gender research, pointed out to me, this figure was about 317 suicide deaths per 100,000 patient-years in the *NEJM* study. That’s quite high. We should be cautious here, because per-100,000 rates are weird when the raw number of events is this low: One fewer suicide would have halved the rate, and one more would have increased it by 50%. For what it’s worth, when I asked the suicide researcher, they responded: “I’d say yes. I agree two suicide deaths in that age group for that sample size is high compared to the general population for sure.”

But it would be unfair to say “Aha, the kids in your group had a high rate of suicide — your treatment doesn’t work.” These were kids who already had some mental health concerns; gender dysphoria itself can be quite distressing and the LGBT population is known, more broadly, to have elevated rates of mental health issues. So we probably

shouldn't expect kids in this study to have the same suicide rate as age-matched peers from the general population. When I raised this point in an email with the suicide researcher, they said that they agreed that "a comparison to suicide rates in another group with mental health struggles would probably be more appropriate than a general population rate."

Unfortunately, we don't have much data here. In a letter published in *Archives of Sexual Behavior*, Biggs had previously calculated that the rate of completed suicide at the Tavistock clinic in England was 13 per 100,000 patient-years, much lower than what was observed in the *NEJM* study. The only other decently apples-to-apples comparison we have at hand here is also mentioned in his letter:

Only one published study has reported suicide fatalities among transgender adolescents. Belgium's pediatric gender clinic provided counseling to 177 youth aged from 12 to 18 years, who had been referred between 2007 and 2016: five of them (2.8%) committed suicide (Van Cauwenberg et al., 2021). The mean age of referral was 15, implying a mean duration of 3 years before transition to an adult clinic, which translates to an annual suicide rate of 942 per 100,000. This is the highest suicide mortality recorded for any transgender population.

So the *NEJM* sample didn't have the sky-high suicide rate of that Belgian cohort, but it's undeniably high.

That doesn't mean the suicides in the *NEJM* cohort were *caused* by the hormones. "Of course we can't attribute those suicides to cross-sex hormones, because we lack a control group," said Biggs in an email. "Likewise, we can't attribute the improvement to the cross-sex hormones!" We'll return to that latter point, but one needn't make a causal argument here to be concerned. One of the most common justifications for why youth gender medicine is worth it, despite the myriad remaining unknowns, is that kids will kill themselves if they don't go on it. Well, here was a sample of kids who had access to it in supposedly high-quality settings, with a lot of support and monitoring, and they still had a very high rate of suicide. How does that not raise questions? The researchers have nothing to say about it, other than noting the number of completed suicides and instances of "ideation."

The issue becomes only more worrisome when you look at the study protocol we spent so much time poring over in the last post and see that kids who had severe psychiatric problems, including suicidality, were excluded from the study at the outset: "Presence of serious psychiatric symptoms (e.g., active hallucinations, thought disorder) that would impair the individual's ability to provide true informed consent or participate in the baseline ACASI [audio computer-assisted self-interviewing]" was an exclusion criterion, as was being "Visibly distraught (e.g., suicidal, homicidal, exhibiting violent behavior) at the time of consent or the baseline ACASI[.]"

So kids could have some degree of suicidality and still participate in the study — researchers *don't* view suicidality as an on-off binary — but kids who were *very* suicidal, or otherwise *very* unwell, were excluded, meaning we actually wouldn't expect

this to be a particularly suicidal cohort going in. And yet, there were still two suicides. That's not good, and should be seen as a red flag that deserves explanation. We have none, because the researchers fail to report on the cohort's overall level of suicidality over time, despite it being part of their core hypothesis. And *The New England Journal of Medicine*, publishing what it knew would be a highly attention-getting study on a hotly controversial issue that is constantly paired with suicide in the public conversation, didn't ask them to.

Most Of The Improvements The Cohort Experienced Were Small

These were the improvements observed over time, all statistically significant, in the variables the researchers reported, according to their statistical model. Remember that the study covers a two-year span:

Appearance congruence: Increase of 0.96 out of 5 points

Positive affect: Increase of 1.6 out of 100 points

Life satisfaction: Increase of 4.64 out of 100 points

Depression: Decrease of 2.54 out of 63 points

Anxiety: Decrease of 2.92 out of 100 points

These numbers are the average changes for the whole group. For the bolded ones, there were statistically significant changes for both sexes. For the non-bolded ones, trans boys but not girls saw benefits. Appearance congruence and positive affect are the only two variables where the researchers were able to report salutary increases in both sexes over the two years of the study. (Appearance congruence is going to get its own section soon, so I'm going to ignore it here.)

Given these differences, it would have been useful for the researchers to lay out plainly what the average changes were for the trans boys versus the trans girls, who were, after all, administered totally different hormones. They don't do that (nor do they offer any speculation as to why two-thirds of their sample were natively female). If you know how to read Table 3 you can sort of reverse engineer some of this information, but it really should be clearer.

The first question you should ask in a situation like this, where you have statistically significant improvements that look small in magnitude, is whether they matter. You can have a statistically significant effect that isn't *clinically significant*, meaning it wouldn't represent noticeable improvement or worsening in the condition being measured. Statisticians argue about effect sizes all the time, but there often isn't an easy answer as to whether a small one is *too* small. I do think sometimes, for super-small effects, it's okay to turn to common sense. Positive affect, as measured by an instrument in the [NIH Toolbox](#), increased by 1.6 points over two years on a 100-point scale. If a researcher touting a treatment for your kid says "This will improve their

score on this self-reported scale 1.6% over two years,” you have every right to be skeptical. I really don't buy the idea that we should care about this or view it as evidence supporting the idea that hormones help kids. (I did check to see if [these documents](#) about the NIH Toolbox had any information about interpreting changes to scores in positive affect or life satisfaction, but they don't appear to.)

What about some of the other results that look small, but not quite so tiny? Should we care about those?

All we can really do is try to look around in the literature and find other comparisons. One of the meatier papers I found on the subject of clinical versus statistical significance was a meta-analysis from the Cochrane Database of Systematic Reviews on “[New generation antidepressants for depression in children and adolescents](#)” (Cochrane is considered one of the best games in town for this sort of careful research evaluation). There, the authors sum up studies that compared these antidepressants to a placebo, and that used something called the Children's Depression Rating Scale-Revised to evaluate symptoms. The CDRS-R has a range of 17 to 113, meaning it's a 97-point scale. The authors describe differences as high as 3.51 as “small and unimportant.” So again, a difference that doesn't appear tiny, per se, might not matter clinically.

More directly applicable to the present discussion, [here's](#) a 2015 paper by researchers examining the concept of minimally important clinical difference, or “the smallest difference in score considered clinically worthwhile by the patient,” as it pertains to the Beck Depression Inventory 2 (BDI-II), which is the item the *NEJM* researchers used to measure depression in their study. The 2015 authors “estimated a MCID of a 17.5% reduction in scores from baseline... The corresponding estimate for individuals with longer duration depression who had not responded to antidepressants was higher at 32%.” In the *NEJM* researcher's statistical model, the patients had a mean baseline score of 15.46 and an average reduction of 2.54. Since 17.5% of 15.46 is about 2.51, if we trust these estimates, the average kid in the *NEJM* study just ever so barely noticed an improvement in their depression symptoms in their two years on hormones. (I couldn't find any research on what constitutes a clinically significant improvement on the anxiety instrument the researchers used, the Revised Children's Manifest Anxiety Scale, where the researchers observed an average improvement of 2.92 out of 100 points.)

Of course, the average kid had only mild depression symptoms to begin with, which makes things more complicated to analyze. Other patients had higher BDI-II scores at baseline, which at least arguably means they'd only view reductions as worthwhile if they were significantly larger.

This is a useful chart from the Supplementary Appendix:

Table S6. Proportions of Youth Scoring in the Clinical Range for Depression and Anxiety at Each Timepoint

	Baseline	6-month	12-month	18-month	24-month
Beck Depression Inventory-II n (%)	<i>n</i> =307	<i>n</i> =281	<i>n</i> =248	<i>n</i> =210	<i>n</i> =219
Minimal Depression	149 (48.5)	152 (54.1)	143 (57.7)	125 (59.5)	126 (57.5)
Mild Depression	53 (17.3)	46 (16.4)	41 (16.5)	25 (11.9)	41 (18.7)
Moderate Depression	57 (18.6)	43 (15.3)	24 (9.7)	30 (14.3)	22 (10)
Severe Depression	48 (15.6)	40 (14.2)	40 (16.1)	30 (14.3)	30 (13.7)
Revised Children's Manifest Anxiety Scale 2	<i>n</i> =308	<i>n</i> =282	<i>n</i> =248	<i>n</i> =209	<i>n</i> =216
<i>M</i> (<i>SD</i>)	60.0 (11.5)	58.6 (11.6)	58.6 (11.3)	56.8 (11.4)	57.4 (12.1)
n (%) in Clinical range (<i>T</i> >60)	181 (58.8)	145 (51.4)	115 (46.4)	90 (43.1)	103 (47.7)

Note. % calculated as valid percent using the *n* for each timepoint as the denominator.

I don't know what to make of this. The researchers tout the fact that a lot of kids moved to lower levels of depression and anxiety over the course of the study, but also acknowledge that a fair number of them remained in the clinical range at follow-up. That's clearly true. This is a genuinely complicated situation to interpret, partially because the baseline numbers are such a mixed bag: It's not fair to harp on the lack of improvement in a kid who wasn't doing that poorly to begin with. This is an issue in some youth gender medicine clinical research, which typically involves cohorts who have been screened for serious mental health problems beforehand: They don't have much room to improve, so the deck is somewhat statistically stacked against researchers seeking to demonstrate mental health improvements. (I guess on the other hand, it's also useful information that the kids in this cohort didn't seem to get significantly worse, for the most part.)

It would have been helpful if the researchers provided more fine-grained information about, say, the average numerical improvement among kids in the moderate or severe ranges on the depression measure. As we've seen previously, reporting instead on the percentage of participants in different clinical categories can obscure a lot of useful information: In the most extreme cases, a one-point drop the patient doesn't even notice can bring them from (say) "moderate" to "mild." It would also be useful to know if the relatively large proportion of kids who didn't provide data at 24 months, which is about a third of them, differed at other time points from the rest of the group, because if the kids with missing 24-month data were on average doing better or worse than their peers in the study, that could seriously skew the results. (I could be missing something, but I think the researchers make this comparison only for the very small number of true dropouts, rather than for kids who technically stayed in the study but didn't provide data on some items at the final observation.)

At the end of the day, it seems hard to deny that a lot of kids stayed unwell despite two years of regular access to a gender clinic and to a medication designed to improve their mental health outcomes. At baseline, 18.6% had moderate depression, and two full years later 10% still did. For severe depression, there was even less improvement: 15.6% to 13.7%. Same general story with anxiety, where 58.8% of the kids scored in the clinical range at baseline and 47.7% did at follow-up.

On the other hand, it's hard to ignore that clearly some kids *did* experience meaningful reductions in depression and/or anxiety and/or other symptoms. Doesn't that, at least, constitute some evidence for the efficacy of hormones?

Unfortunately...

It's Impossible To Attribute The Improvements Observed In This Study To Hormones Rather Than Other Forms Of Treatment That Took Place At These Clinics

This is a longitudinal study without a comparison group. As any graduate of an AP Statistics course will tell you, this makes it much harder to claim that any particular influence is responsible for any changes that are observed over time.

If you track *two* otherwise similar groups over two years, and you give one of them medicine and the other a placebo, and you observe differences between the groups, you might then be able to begin to make some reasonably confident causal inferences about the effects of the medicine, though *how* confident depends on a host of factors. But with only one group, this is a notorious statistical problem, even in relatively simple situations. If I give you flu medication, and your flu symptoms have improved significantly five days later, does that mean the meds worked? Maybe. But people also tend to get better over time. Scores on certain scales tend to revert to the mean if the first observation is quite high or very low. And so on. Again, these are really basic statistical principles — this isn't nitpicking.

This *NEJM* study is far more complicated than a two-week flu study, though. And we actually have some pretty good reasons to suspect other factors may have contributed to the (mostly small) improvements observed by the researchers.

As they note early in their paper, “All participating clinics employ a multidisciplinary team that includes medical and mental health providers and that collaboratively determines whether gender dysphoria is present and whether gender-affirming medical care is appropriate. For minors, parental consent is required to initiate medical treatment. Publications by individual study teams provide details on site-specific approaches to care.”

That last sentence cites four papers, and if you read those papers, you'll see some mentions of therapy and medication for patients experiencing mental health duress above and beyond their gender issues. To take one example, the team at the Gender Identity & Sex Development Program at Lurie Children's Hospital in Chicago writes that “In cases when both evidence-based therapy and psychopharmacotherapy are indicated, a psychologist and the psychiatrist may comprise a patient's treatment team — the psychologist serving as a primary therapist and the psychiatrist offering psychotropic medication management.”

Every kid in this study was seen at one of these multidisciplinary clinics. So it stands to reason that the ones with serious anxiety and depression issues were likely provided with access to psychotherapy, medication, or both. It also stands to reason that the worse a kid's mental health symptoms were at baseline, the more likely they were to have been provided with one of these interventions, *and* the more room they had to improve. This means that even when it comes to the subset of kids whose improvements were sizable, we have to ask: Did their symptoms abate because of the hormones, the medication, or the therapy? Or was it some combination of all three?

There's *no* way to know. This makes it effectively impossible to interpret these results fully. And it's unfortunate, because I believe the researchers could have potentially accounted for this in their statistical models since surely they had access to patient records. One of the questions I sent them was on this very subject, but as I mentioned in my last post, they aren't doing any interviews or responding to queries.

Despite all this, the researchers confidently proclaim in their abstract that “[Gender-affirming hormones] improved appearance congruence and psychosocial functioning.” This is straightforwardly causal language, but their methodology doesn't come close to warranting it. Moreover, they don't even mention this potential confound, which, again, really is the sort of thing you would learn during the first year of a college-level introductory stats course.

Imagine encountering this question on the midterm of such a course:

A group of kids with mental health problems gets access to psychological counseling, psychiatric medication, and Treatment X for two years. At the end of this period they are doing significantly better. Is this because of

- a) the counseling
- b) the medication
- c) Treatment X
- d) without more information, it's impossible to know

One hundred out of one hundred stats professors would tell you the correct answer is (d). (An accompanying *NEJM* commentary on the paper by Annelou de Vries and Sabine Hannema mentions the therapy issue but not the medication issue.)

There's one other potential problem with pinning the improvements noticed in this study on the powers of gender-affirming hormones, per se. It was pointed out to me by a depression researcher with whom I exchanged some emails about this paper. He initially said I could quote him by name but changed his tune, saying the subject was too fraught. When I asked if it would be okay to describe him as “a depression researcher,” he responded, “Sure, I'd even be ok with ‘a depression researcher who's too chickenshit to be named...’ 😊.” (I include this because I thought it was funny, not because I think he's chickenshit! I don't blame him at all.)

He pointed out that the researchers “utterly ignore the obvious point that testosterone has large mood-elevating, anti-anxiety, and antidepressant effects!” I’d heard others raise this point in the past, and it poses another genuine challenge for youth gender medicine researchers. There is **indeed some evidence** that testosterone has mood-improving effects, and this raises the possibility that it can cause improvements in mental-health symptoms that don’t have to do with treating gender dysphoria per se. (It should be clear that the studies I’m referencing and linking to generally involve natal males rather than natal females, adults rather than juveniles, and sometimes cover periods of time significantly shorter than two years. So we can’t state definitively that T has similar effects on young natal females than it does on older natal males, but it’s definitely a possibility, and there’s some anecdotal evidence of this among trans boys and men.)

Let’s again be maximally generous and set aside this possibility. Summing up the rest of this section, even if we put a thumb on the scale and zoom in *only* on the kids who got a lot better in this study, ignoring the small average effect sizes, the suicides, the truly disappointing results for trans girls, and what appears to be a veritable carnival of variable cherry-picking, we simply have no way of knowing — full stop — whether they improved because of hormones, therapy, medication, or a combination of the three.

I am not good at stats, but I don’t need to be to make this point. It is not esoteric. This is a very basic issue, well known in the social sciences. It seriously undermines our ability to determine whether the kids in this study benefited from going on hormones, and it reflects poorly on *The New England Journal of Medicine’s* decision to allow the researchers to use such strong, straightforward causal language to describe their results.

The One Bigger Improvement Was In A Variable That Might Not Mean All That Much

So far we’ve been arguably talking about peanuts, as far as average effect sizes are concerned. The authors do highlight one more impressive-seeming finding, though:

Increasing appearance congruence is a primary goal of GAH, and we observed appearance congruence improve over 2 years of treatment. This was a moderate effect, and the strongest effect observed across our outcomes, consistent with the effect seen in research involving other samples, which has noted large effects of GAH on body image and small-to-moderate effects on mental health. Appearance congruence was also associated with each psychosocial outcome assessed at baseline and during the follow-up period, such that increases in appearance congruence were associated with decreases in depression and anxiety symptoms and increases in positive affect and life satisfaction. These findings suggest that appearance congruence is a candidate mechanism by which [gender-affirming hormones] influences psychosocial functioning. [footnote omitted]

Specifically, over two years the kids in the study experienced about a one-point improvement on the five-point appearance congruence scale of the Transgender Congruence Scale.

The TCS is a 10- or 12-item instrument with the following items, as summarized in a table from a 2021 study in *Sexuality Research and Social Policy*, which was the most recent one I could find that had been conducted on its psychometric properties:

Table 3 Factor loadings for original TCS and TCS-10 reduced models

Items by original factors	Item #	Original λ	TCS-10 λ	Factor in TCS-10 model/reason removed
Appearance Congruence (Kozee et al., 2012)				
My outward appearance represented my gender identity	1	.493	.481	Appearance Congruence
I experienced a sense of unity between my gender identity and my body	2	.498	.447	Appearance Congruence
My physical appearance adequately expressed my gender identity	3	.561	.604	Appearance Congruence
I was generally comfortable with how others perceived my gender identity when they look at me	4	.624	.643	Appearance Congruence
My physical body represented my gender identity	5	.559	–	Removed due to high covariance
The way my body currently looks did not represent my gender identity (Reversed)	6	.276	.269	Appearance Congruence
I was happy with the way my appearance expressed my gender identity	7	.457	.491	Appearance Congruence
I did not feel that my appearance reflects my gender identity (Reversed)	8	.280	.316	Appearance Congruence
I felt that my mind and body were consistent with one another	9	.436	.398	Appearance Congruence
Gender Identity Acceptance (Kozee et al., 2012)				
I was not proud of my gender identity (Reversed)	10	.391	–	Removed due to high covariance
I was happy that I have the gender identity that I do	11	.555	.494	Gender Identity Acceptance
I had accepted my gender identity	12	.537	.561	Gender Identity Acceptance

This past fall I wrote a piece about the problem of researchers in this field touting impressive-seeming findings by relying on instruments that might not mean much. In that case, I discussed adolescent top surgery research where the headline finding was that kids' scores "improved," after receiving double mastectomies, on scales that seemed to consist mostly of items asking them whether they presently had breasts. That is, if you ask someone with gender dysphoria to rate their agreement with "I worry that people are looking at my chest" — an actual item from the scale in question — before and after they have surgery, it would be shocking if their score *didn't* improve. But that obviously doesn't tell us much about whether double mastectomies "work" in the longer-term sense we would want a major surgery to work. The paper in question didn't include more substantive items on the kids' mental health.

I think there's a version of that going on here. To be fair, the authors of the *NEJM* paper also report on some more common and better-validated measures, including ones addressing anxiety and depression, but as we've seen, those changes were small, of questionable clinical importance, and didn't apply to the male-to-female transitioners. Improvements in appearance congruence were experienced by both sexes, and it is the *only* decent-sized effect the researchers uncovered, as they themselves note.

But there's a case to be made that the game is somewhat rigged here. Take items like "My outward appearance represented my gender identity" and "My physical appearance adequately expressed my gender identity." For one thing, these items are so similar that I'm surprised they're both on the scale — they seem truly redundant, and it seems almost impossible that if one goes up or down, the other isn't going to

follow right along with it, which might artificially inflate observed changes in respondents' scores. (In theory, when a scale is first validated, someone checks for these sorts of issues, and I'm not going to claim to have looked deeply into that process here. Though it's worth noting that in the 2021 paper, the authors note one item on the appearance congruence subscale was removed due to its high covariance with others.)

But more importantly, it seems almost impossible to imagine how someone's scores on an item like this *wouldn't* "improve" as the physical changes of hormones took hold and brought their body in line with who they felt they were on the inside. I mean, I don't want to discount this finding entirely — it would certainly be bad news if it didn't improve, because it might suggest their gender identity or transition goals had changed mid-treatment (which wouldn't be ideal) — but I'm just not sure how impressed we should be by this.

That's doubly true when you notice that the appearance congruence scale doesn't appear to really correlate with other, more robust measures of well-being, anyway. Or at least that's what the authors of the 2021 paper found:

Table 4 Correlations between *TCS-10 Total* and subscales with gender and well-being constructs including means, standard deviations, and range

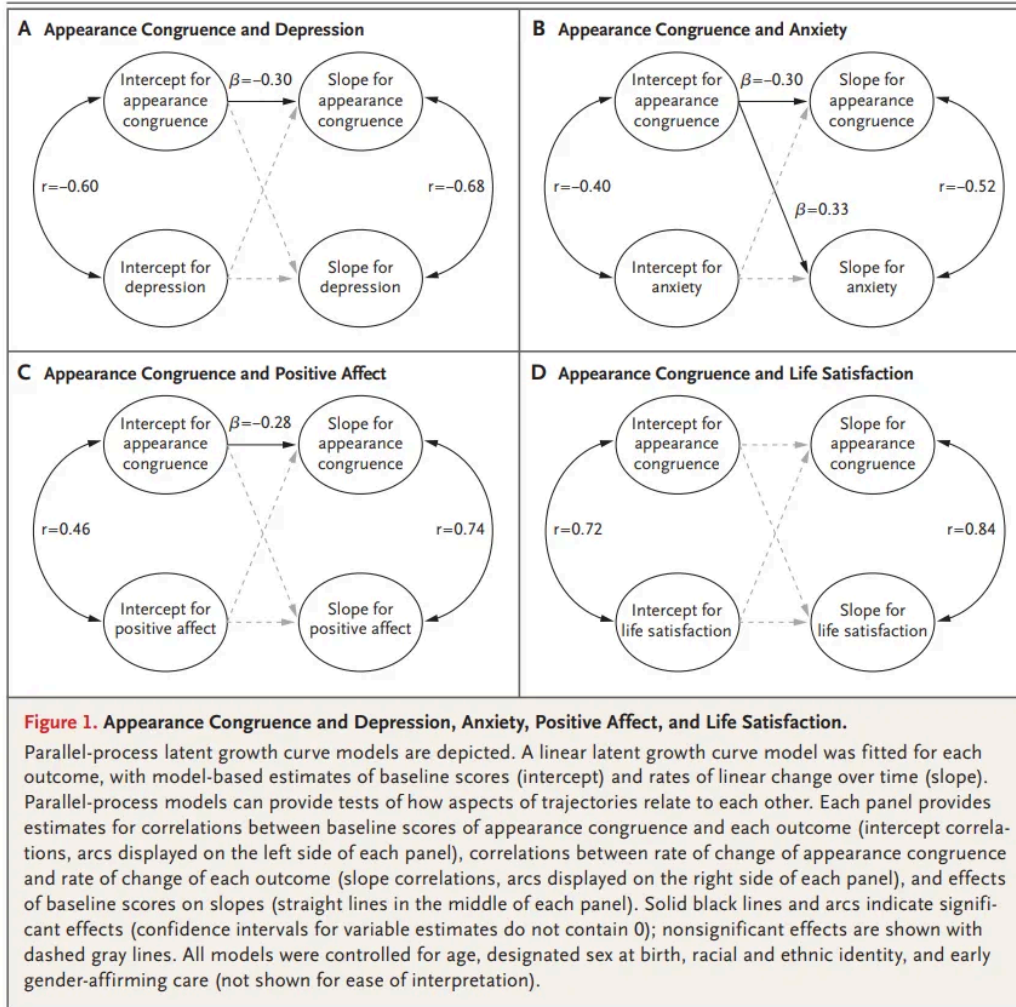
	Total	AC	GIA	M	SD	Range
TCS-10 Total	—	.96***	.50***	3.24	.60	1.70–5.00
- Appearance Congruence (AC)	.96***	—	.22**	3.18	.67	1.00–5.00
- Gender Identity Acceptance (GIA)	.50***	.22**	—	2.34	.61	1.00–3.33
TC ³	.58***	.54***	.33***	56.28	9.18	28.00–89.00
GPSQ	-.23**	-.18*	-.26***	41.83	8.20	15.00–62.00
GMSR						
- Discrimination	-.19**	-.11	-.29***	3.63	1.69	0.00–5.00
- Rejection	-.12	-.06	-.21**	4.15	2.06	0.00–6.00
- Victimization	-.04	.05	-.26***	3.99	2.26	0.00–6.00
- Non-affirmation	-.25***	-.28***	.00	14.14	4.69	0.00–24.00
- Internalized transphobia	-.18**	-.09	-.34***	16.77	6.75	0.00–30.00
- Pride	.36***	.26***	.44***	18.66	5.84	4.00–32.00
- Negative expectations	-.02	-.03	.02	20.25	6.52	0.00–36.00
- Nondisclosure of gender identity	.09	.13	-.07	10.98	4.06	0.00–20.00
PHQ-9	-.11	-.05	-.22**	13.01	5.81	0.00–27.00
GAD-7	-.02	.02	-.11	10.74	4.60	0.00–21.00
PANAS						
- Positive	.31***	.26***	.25***	28.90	6.23	11.00–45.00
- Negative	-.03	.03	-.19**	26.17	7.42	9.00–40.00
SWLS	.43***	.43***	.17*	21.59	5.77	6.00–35.00

Note: *TCS*, Transgender Congruence Scale; *TC³*, Trans Collaborations Clinical Check-in; *GPSQ*, Gender Preoccupation and Stability Questionnaire; *GMSR*, Gender Minority Stress and Resilience; *PHQ-9*, Patient Health Questionnaire (9-item); *GAD-7*, Generalized Anxiety Disorder (7-item); *PANAS*, Positive and Negative Affect Scale; *SWLS*, Satisfaction with Life Scale. *N*'s range from 202 to 208 due to missing data

*** $p < .001$, ** $p < .01$, * $p < .05$

Don't worry if you can't interpret this. The point is that this measure only inconsistently correlates with other, better-established ones. Without getting too into the weeds, the authors do note that this conflicts with [previous research](#) into the TCS and its subscales that found it *did* correlate, pretty strongly, with other items. But it seems like an open question whether changes in the TCS or its subscales matter all that much, clinically — a question worth investigating further, but an unanswered one. If this is the biggest improvement you note in your big *NEJM* study, you should be asking yourself some questions.

The researchers argue that they did some other statistical work supporting the idea that appearance congruence might be particularly important. They include this fancy-looking visual:



One of my go-to folks for stats stuff that is above my head is Stuart Ritchie, a research psychologist and the author of the great book *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth*, as well as his own Substack newsletter of the same name that was recently acquired by the UK's *i* newspaper.

I sent him this chart and the authors' claims about their latent growth curve modeling and what it showed about the importance of appearance congruence. As luck would have it, Ritchie had published research employing similar statistical techniques. "To me it makes sense that the appearance congruence would change alongside the other mental health variables - but that doesn't say anything about whether changing appearance congruence would change mental health in a causal sense," Ritchie wrote in an email. "The causality could easily have just gone the other way (people who felt better, mentally speaking, tended to worry less about their appearance)." Ritchie summed up this part of the *NEJM* paper as "a fancy way of messing around with correlations, nothing causal, kind of interesting but not a clincher of any argument."

The anonymous depression researcher made the exact same point, independently, in an email. “The latent growth curve models are not terribly persuasive,” he wrote. “Sure, they suggest that changes in appearance congruence are correlated with changes in some other psych variables, but there’s no proof of *causality* there, and we certainly know that self-reported appearance measures are heavily mood-state-dependent – so an increase in depression, say, would almost certainly cause a decrease in appearance congruence.”

Ritchie and the depression researcher also both independently noted the high number of statistically insignificant links (the dashed lines) in the above diagram. There doesn’t appear to be a lot there.

The Researchers Don’t Even Consider The Possibility These Treatments Are Less Effective Than They Thought — Their Only Answer Is “More Hormones”

There’s a subtly revealing part of this paper where the researchers attempt to grapple with the fact that the natal males in the study hardly showed any improvements during their first two years on hormones:

Given that some key estrogen-mediated phenotypic changes can take between 2 and 5 years to reach their maximum effect (e.g., breast growth), we speculate that a longer follow-up period may be necessary to see an effect on depression, anxiety, and life satisfaction. Furthermore, changes that are associated with an endogenous testosterone-mediated puberty (e.g., deeper voice) may be more pronounced and observable than those associated with an endogenous estrogen-mediated puberty. Thus, we hypothesize that observed differences in depression, anxiety, and life satisfaction among youth designated female at birth as compared with those designated male at birth may be related to differential experiences of gender minority stress, which could arise from differences in societal acceptance of transfeminine (i.e., persons designated male at birth who identity [sic] along the feminine spectrum) as compared with transmasculine persons. Indeed, gender minority stress is consistently associated with more negative mental health outcomes, and research suggests that transfeminine youth may experience more minority stress than transmasculine youth. [citations omitted]

Two things on this: First, I don’t think this really jibes with the fact that the improvements in appearance congruence were statistically equal between natal males and females. Of course appearance congruence is only one aspect of a successful transition, but you would think that if all these other obstacles were getting in the way of trans girls feeling better after two years on hormones, it would show up in the variable that most closely tracks the physical effects of those hormones.

More importantly, science is supposed to be open-minded. If you’re evaluating a new treatment, you’re supposed to attend to the possibility that it doesn’t work as intended.

It could be everything the authors are saying in this excerpt is true, but it's impossible to ignore the chain of events here: They put a cohort of kids on hormones, two of them died from suicide, and the natal males appear to have experienced no measurable improvements other than a truly tiny one on a positive affect scale and a questionably important one on a rather tautological appearance congruence scale. Their response is to say... maybe the kids just need to be on hormones longer. There isn't even a moment of pause or reflection or uncertainty. It's full steam ahead.

The fact is that some members of this team aren't just clinicians and researchers — they're also steadfast advocates for these treatments. They *strongly* believe that these treatments help trans kids, and they benefit materially from administering them and from participating in the public debate about them. This situation, in which strong advocates with clearly stated preexisting views are producing what is supposed to be top-tier evidence (it's *The New England Journal of Medicine*, after all), shouldn't concern us a *little*?

It's probably inevitable that in some cases advocates for a treatment are also going to research that treatment. But we should at least acknowledge the potential pitfalls here. When Cochrane or the UK's National Institute for Health and Care Excellence (NICE) publish scientific evidence, there's the expectation that the bureaucrat-nerds responsible for producing it are entering the process without strong priors, and are reasonably invested in examining the evidence in a fair and evenhanded manner. In fact, part of the reason NICE's [damning reviews](#) of the evidence for giving puberty blockers to kids and hormones to adolescents marked a major turning point in this discussion is *because* it is such a trusted institution.

If Cochrane published an evidence review showing that one particular antidepressant outshined others by a wide margin, and it was subsequently revealed that a coauthor on that review had a husband who worked for the pharmaceutical company that produced the antidepressant in question — and that this relationship hadn't been disclosed — that would immediately cast a dark shadow across the article. It would be considered a conflict of interest.

Why should *none* of that logic apply here? A single *NEJM* article isn't the same as a comprehensive evidence review, sure, but why should we completely ignore the reality, which is that since humans are humans, a team of researchers deeply invested in a treatment might be less capable of carefully, dispassionately evaluating the evidence for it?

Maybe there's a reason why none of that logic should apply here at all. But if there is, I can't come up with it. Especially in the continuing absence of any sorts of systematic reviews of youth gender medicine in the US — an important point Moti Gorin made [here](#) — each study on this subject is going to take on an outsized importance. Which is all the more reason to demand that the researchers behind these studies adhere to the highest standards of rigor and transparency.

I don't think that happened here, and it isn't the first time strong advocates for youth gender transition have produced questionable research about it. As I noted in Part 1, in this case it might not be fair to lay all the question marks at the feet of the researchers themselves; some of them may be the result of the *NEJM*'s editorial input. But still: We're now many years and many hundreds of thousands of dollars into this research effort. The researchers have published what was supposed to be one of their blockbuster studies, and it is missing absolutely crucial evidence we need in order to evaluate these treatments. All these years later, we don't even know whether the cohort experienced reduced suicidality or gender dysphoria. What we *do* know is not encouraging: a high suicide rate, tiny-to-small improvements except on a questionable measure, *no* improvements on most measures for the trans girls.

Again: It could be that if we properly adjusted for all the missing, unreported variables, the researchers didn't really find anything at all. It's frustrating that we're still in a position where that's a live possibility — more of these questions should have been answered by now.

Questions? Comments? Unfounded claims? I'm at singalmindeed@gmail.com or on Twitter at [@jessesingal](https://twitter.com/jessesingal).



166 Likes

Discussion about this post

Comments

Restacks



Write a comment...



Frederick R Prete 7 Feb 2023

...

As a Biological Psychologist and research scientist, I agree with your (detailed) analysis. However, in some sense it's beside the point (with which, I think, you probably agree). People see what they want to see in data, especially data that deal with fuzzy psychological concepts, or which can be spun to support an ideological agenda.

However, probably the most important issue, from my perspective, is the fact that you cannot simply manipulate anatomy and physiology to change a person into something they would like to be (or you would like them to be). Biology doesn't work like that. Further, we have known the deleterious effects of cross sex hormones and blockers for over half a century (ever since the drugs were first synthesized). It amazes me that people are now "debating" their effects. Except in the rarest of instances will traumatic hormonal manipulation help someone who is suffering from a psychological or psychosocial condition. I discussed some of that in "There is "Biological Evidence for Gender Identity..." but it's not what you think."

<https://everythingisbiology.substack.com/p/there-is-biological-evidence-for>

Interestingly, much of the back-and-forth about "gender identity," its associated issues and their putative amelioration is driven by misconceptions about biological causation, biological malleability, and fundamental misunderstandings of neuroscience on both sides of the debate. In some ways, it's

become like the debate over abortion; misconceived by people on both sides of the issue, and subject to a perpetual attempt to win the argument by presenting progressively larger bodies of "facts".... with no one making a fair and honest assessment of their "opponents" point of view.

To be clear, I agree with your analysis. I'm just not sure if it's going to make much of a difference until people decide to more carefully — and objectively — consider the fundamental biological and psychological aspects of this issue. Everything is biology, after all.

Thanks for a great read, Frederick

♡ LIKE (30) 💬 REPLY

📤 SHARE

4 replies



John Caramel 7 Feb 2023 *Edited*



Hi Jesse,

I only ever seem to comment or tweet you when I want to disagree so first of all I have to state what great work you do and how much I appreciate it.

But with every write up like this that I read, I less understand your position on supporting hormones for *some* children as a better solution than, say, leeches.

Have you written about the positive research or the persuasive arguments that have led to this position? About which are the cases you can definitively make such a prescription for? If not, would you? Please.

Thanks again

♡ LIKE (21) 💬 REPLY

📤 SHARE

17 replies

58 more comments...