CSPI

About Those Baby Brainwaves: Why "Policy Relevant" Social Science is Mostly a Fraud

Jordan Lasker*

Center for the Study of Partisanship and Ideology

Report No. 5

4/4/2022

To read the executive summary, click here.

Code is available for all original analyses.[1]

* PhD Student at Texas Tech University

CSPI

Table of Contents

Table of Contents **2**

Summary **3**

Policy, Politics, and Social Science **3**

Cash and Baby Brainwaves **4**

Problems With the Literature Linking EEGs to Psychological Outcomes **6**

Small Sample Sizes 6

Inconsistent Foundations 10

Confounded Associations 11

Unrepresentative Samples 11

Multiple Comparisons 12

Problems With the Study Itself **14**

The Unregistered Tests May Have Contradicted the Authors 14

The Study was Reliant on Typically Unimportant Analytic Decisions 15

Other EEG Analyses 16

Researcher Degrees of Freedom 17

Did the Intervention Affect Cognitive Development? 18

Addressing Attrition and Missingness 19

Did the Study Show Parents Spent the Money Well? 19

Should We Expect Effects to Grow? 20

It's a "Contribution" 20

Most Interventions Have Small to Nil Effects **21**

Educational Interventions 21

Cash Transfer Effects 22

Nutritional Interventions 23

Deworming 23

Brain Training 23

Will the Effects be Larger with Greater Cash Transfers? 24

Conclusion **25**

Summary

1. A recent study claimed small cash transfers to the parents of newborns improved their babies' brain activity. The study was lauded in the media and by D.C. policymakers who argued its results supported redistributive policies, most notably the child tax credit.
2. Closer inspection of the study reveals its claims to be wildly overstated and its methodology to be suspect, as is common for policy relevant social science.
3. Authors Troller-Renfree et al. committed numerous bad research practices, including:
    - Deviating from their analysis plan and justifying it by referencing studies that did not support and even contradicted their results.
    - Highlighting results obtained through typically unimportant methodological decisions, which weren't preregistered.
    - Ignoring that larger and more prolonged interventions typically generate smaller effects than they found, meaning their results were likely due to chance or error.
4. Looking beyond the study in question, the theoretical and empirical basis it claims to build upon is largely a house of sand. The evidence that is purported to show a link between brain waves and policy outcomes is extremely weak, and most interventions seeking to improve cognitive ability in children have small to nil effects.
5. Despite its inconclusiveness, the study was portrayed as relevant to child tax credit policy, vigorously promoted in media outlets like *Vox* and *The New York Times*, and championed by think tanks and policymakers.
6. These issues are not unique to the baby brainwaves study. Social scientists frequently engage in questionable research practices and exaggerate the strength and implications of their findings. Policymakers capitalize on low-quality social science research to justify their own agendas. This is particularly true when it comes to research that claims to improve cognitive and behavioral outcomes.
7. Thus, we ought to be skeptical of social scientists' ability to reliably inform public policy, and of policymakers' ability to evaluate social-scientific research objectively, particularly when it is "policy relevant."

Policy, Politics, and Social Science

Social science often informs, guides, and justifies public policy. Theories and empirical findings in economics, sociology, psychology, and political science can inspire and legitimize new programs such as behavioral interventions to improve public health, organizational and administrative schemes to boost diversity, and efforts to combat inequality by redistributing income. Given the credibility of social science with policymakers and the public, one might be inclined to believe that its findings and recommendations are generally useful and accurate. Yet much of the social science informing policy is low quality, and policymakers often benefit from uncritically citing second-rate studies that support their preferred policies.

Political actors aiming to promote a certain outcome frequently use the language and methods of social science to provide evidence that policies they want are effective. Some, like Austrian economists or advocates of Modern Monetary Theory, wield theories as their weapons in the war over policy. Others stick to empirical evidence, which can often be manipulated into supporting any number of narratives. The former typically contort facts to fit with theories and present the beneficial conclusions of their theories as a result, thus justifying policies that follow from their theories. The latter overfit, distort, omit, and sculpt evidence to fit a narrative,

however convoluted, and present their work as an affirmation of said narrative and their expectations of it. While both types of actors do a disservice to science, this report is focused on the latter school of empirical policy-centric social science. It focuses on one particular study, the research that it supposedly builds upon, and literatures related to it to demonstrate how far away the social sciences are from being able to reliably inform policy debates.

To give an example, consider the concept of food deserts: areas in which people are believed to have poor access to healthy foods. They have been the subjects of intensive investigation; studies correlating access measures, poverty, obesity, and various other outcomes all seem to confirm that food deserts cause poor health and drive disparities between rich and poor. Yet, the research basis for food deserts has been almost entirely uninformative. It has largely consisted of research designs supporting the conclusion that food deserts are related to bad outcomes, correlating those outcomes with variables like poverty, obesity, and the proximity of supermarkets. However, there are methods to assess the causal impacts of food deserts, not just their associations. The establishment of government-subsidized supermarkets, supermarkets opening due to tax incentives, or a variety of small food store interventions have all been fruitfully leveraged to gauge the effects of changes in the food environment on all the important correlates of living in a food desert.[2] They have, at best, delivered effects on ill-defined measures of 'knowledge,' like which foods are fattier, and modestly increased the amounts of fruit and vegetables people eat. The 'improvements' seen in these studies are not clearly related to health at all, and studies have *not* supported improvements in objectively measured health, as indicated by obesity rates or other health measures.[3] Taken seriously, the causal studies suggest that socioeconomic disparities transcend differences in access to healthy food.

The academics, policymakers, and activists who in the past promoted food desert studies that were purely associational never produced sufficient evidence for their policy guidance. They contributed to a now-widespread belief that disparities in access to healthy food drive socioeconomic differences in health when, to date, the evidence to that effect is extraordinarily weak at best. This naïve analysis has been used to justify many public and corporate policies, including the Obama administration's nationwide program to end food deserts, Lyft's Grocery Access Program in D.C., Walmart opening more than 200 stores in areas with food deserts, and the National Institute of Health issuing millions of dollars in grants to research food deserts.[4] And yet, the best evidence suggests that these kinds of policies won't improve health, because food deserts aren't causing their inhabitants to be unhealthy.

There are many other examples of weak social science being used to questionable ends: ineffective school management interventions foisted on hundreds of Indian schools, decades of wasteful "Drug Abuse Resistance Education," and lucrative grants for fraudulent nutrition science, not to mention the plethora of ineffective "nudging" policies and practices inspired by social psychology and behavioral science.[5] It is intuitive that research will influence popular political will, and, with qualifications, it probably *should*. But when research is weak or, in the worst cases, based on fraud, we might suffer from its political effects. A bad research paper can hurt an academic's career, but a bad policy can drastically harm individuals, communities, and even countries. Thus, those with the relevant technical training – scientists and hard-nosed policy analysts, for example – should be openly and vocally critical of bad social science, especially when it is policy relevant.

Cash and Baby Brainwaves

On January 24th, 2022, *The New York Times* published a "Breaking News" piece entitled "Cash Aid to Poor Mothers Increases Brain Activity in Babies, Study Finds."[6] The article

discussed Troller-Renfree et al. (2022), a study published in the *Proceedings of the National Academy of Science* (PNAS) that purported to show a monthly subsidy of $333 yielded considerable improvements in children's brainwave activity.[7] *Vox* also covered the story, opining that "we can be reasonably confident the cash [parents received] is a primary cause of these changes in babies' brains. And we can be reasonably confident it will be a causative factor in whatever future outcomes the… researchers find."[8] They weren't alone: think tanks, medical news aggregators, and other mainstream outlets like *NBC* and *Forbes* all praised the study and the alleged psychological benefits of brainwave improvements.[9]

Laudatory coverage was accompanied by claims that the study had special relevance to expanding child tax credits. The authors even put out a pro-child tax credit press release on their website: "This study's findings on infant brain activity… really speak to how anti-poverty policies – including the types of expanded child tax credits being debated in the U.S. – can and should be viewed as investments in children."[10]

High praise and lofty implications aside, a closer examination of the study reveals its results to be dubious and its methods severely flawed. While the authors' public statements portrayed the study as strong, reliable, and policy relevant, none of these claims hold up to scrutiny. The study and its portrayal stand as excellent examples of what can go wrong when political desire supersedes scientific reasoning, and why we ought to be skeptical of social scientists' ability to reliably inform public policy with their research.

Troller-Renfree et al. (2022) was an analysis of an ongoing project known as the Baby's First Years study. It comprised a total sample of 1,000 low-income (approximately $20,000 per year) mothers and their children, randomly divided into a treatment group in which parents were given monthly subsidies of $333, and a control group that received $20 a month. The cash benefits were considerable in the treatment group, boosting their incomes by around 20% (compared to 1% in the control group). The point of the project was to "provide the first definitive understanding of the extent to which income plays a causal role in determining early child cognitive, socio-emotional and brain development among low-income families."

The study looked at differences in babies' brainwaves between control and treatment groups. Brainwaves are fluctuating electrical pulses produced by the actions of neurons in the brain that can be assessed by electroencephalography (EEG). Brainwaves have been used to measure neural activity and, as the authors described, are usually measured along two dimensions: frequency and power. Frequency is the spectrum along which activity is measured, in Hz, and power is the amount of activity at certain frequencies. Power is often subdivided into absolute and relative power, with the difference being the amount of activity at a given frequency versus the absolute power as a fraction of absolute power across a certain range of frequencies. These frequencies are subdivided into ranges or bands which are given Greek-lettered names, like delta (0.5 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 13 Hz), beta (13 – 35 Hz), and gamma (>35 Hz), giving us terms like "alpha power" or "delta power" to denote power within those ranges. The ranges for these specific brainwaves vary slightly by source, but the order I've listed them in places them in their consistent relative positions from lowest to highest frequency. Power in some of them has been related to a variety of outcomes.

To establish the importance of brainwaves for their analysis, Troller-Renfree et al. cited several studies that found relationships between EEG power and important cognitive and developmental outcomes. They then described the treatment effect of cash gifts on specific brainwave bands using regressions to control for several variables in the part of their study that had EEG data. This is how they worded their results:

> [I]nfants whose mothers were randomized at the time of their birth to receive a large monthly unconditional cash transfer showed greater mid- to high-frequency absolute EEG power in the alpha-, beta-, and gamma-bands, compared with infants whose mothers were randomized to receive a nominal monthly unconditional cash transfer.[11]

Several academics and public intellectuals pointed out flaws in Troller-Renfree et al.'s study shortly after it was released. Psychologist Stuart Ritchie noted the article's potentially-dubious peer review status, the disconnect between observations made by the study's authors and outcomes that matter, the potential for fade-out effects, deviations from preregistered methodology, and the nonsignificance of the reported effects of cash transfers on any of the study's main outcomes.[12] Statistician Andrew Gelman found that the significance of results was not robust and that the graphs produced by the authors could be replicated even if the children in the sample were randomized into artificial treatment and control groups.[13] Finally, psychiatrist Scott Alexander offered further commentary on the article to the effect that the results seemed unlikely for various reasons, alongside a summary of what other people had said.[14]

All three critical commentators agreed: Not only did the study authors make sloppy methodological choices, but they also went out of their way to portray their findings as more policy relevant than they were. Moreover, even if their findings were wholly accurate, which is highly unlikely, their policy suggestions would not follow from their results.

Problems With the Literature Linking EEGs to Psychological Outcomes

For the results of the study to be important when taken at face value, changes in EEG power must either be causally linked to changes in their psychological correlates – such as IQ or language skills – or they must reflect a common cause for change in both. In either case, the results must transfer to the phenomena that EEG power has been linked to in the past. The study's authors cited several pieces of research that linked EEG power to their desired cognitive outcomes. These studies included Benasich et al. (2008), Gou, Choudhury and Benasich (2011), Brito et al. (2016), Maguire and Schneider (2019), Williams et al. (2012), and Brito et al. (2019).[15] Cumulatively, Troller-Renfree et al. argued that these studies linked absolute power in mid-to-high frequency bands to linguistic, cognitive, and social-emotional development, while simultaneously arguing that low-frequency band power was linked to behavioral and learning issues.

These studies suffered from several issues, five of which I will address. First, all had small sample sizes ranging from 13 to 129 participants. Second, the studies they cited were frequently inconsistent, with results from different studies pointing in different directions. Third, these studies' associations with socioeconomic status were almost all confounded with other characteristics like genes and family environment.[16] Fourth, the study samples were often highly unrepresentative of the general population. Finally, many did not address the crucial issue of multiple comparisons, which means it is likely their policy relevant results were due to chance or p-hacking.[17]

Small Sample Sizes

Small sample sizes are a serious problem in scientific research because they lack statistical power: the probability that a statistical test detects a real effect of a given size at some level of significance. If I have a sample size of 20 and I need to find a correlation between X and

Y that has a *p*-value below 0.05, the lowest significant correlation would be Pearson's $r = 0.44$. If I quintuple my sample size, the lowest significant correlation is 0.20. This does not mean that a small sample cannot yield small correlations, only that, when it does, they will not be significant.

When small samples are combined with reviewers and editors being biased towards results that are statistically significant, effect sizes will tend to be exaggerated. Think of it this way: to get published, I need a result with a *p*-value below 0.05. In terms of the typical level of statistical power people in this area seek (80%) and the typical *p*-value cutoff they use (0.05), I would have 80% power to detect a correlation of $r = 0.58$ with a sample size of 20, and 80% power to detect a correlation of 0.28 if I increased my sample size to 100. If I'm stuck with a small study, I must have a big effect or else what I find will not be significant. If I get an effect, it is probably going to be much larger than it would if I had more power to detect it.[18]

We can think of this another way, by comparing the mean significant effect versus the actual effect in a comparison of two group means. At 100% power, the mean significant effect is the same size as the actual effect. However, at 80% power, the mean significant effect is 20% larger than the actual effect, and at 30% power, the mean significant effect is 70% larger than the actual one. The extent of statistical exaggeration is almost twice as large for correlations.

So, how were the studies that purported to link EEG activity to cognitive outcomes? Their relevant descriptive statistics are in Table 1. Though I am not discussing them in this section, the studies linking EEG activity to SES are covered in Table 2.

Table 1. Study Descriptives for Cited Cognitive Ability-EEG Studies[19]

Study	Outcome	Multiple Comparisons	Sample Size *	Effect Size **	Critical *r* ***	Power †
Benasich et al. (2008)	Associations with Absolute Gamma Power	Not corrected for	At most, 63	0.52 - 0.75	0.25	0.12
Harmony et al. (1990)	Associations with Absolute and Relative Power (Various)	Not corrected for	At most, 81	0.52 – 0.93	0.22	0.14
Gou, Choudhury & Benasich (2011)	Associations with Absolute Gamma Power	Not corrected for	At most, 17	0.51 – 0.72	0.49	0.06
Williams et al. (2012)	Associations with Absolute Beta Power	*p*-value cutoff set to 0.006	13	0.77‡	0.72‹	0.01
Brito et al. (2016)	Associations with Absolute Gamma Power	Not corrected for	At most, 64	0.28 – 0.32	0.25	0.12
Brito et al. (2019)	Associations with Absolute Power (Various)	Vague: "running FDR to correct"	At most, 129	ns	0.17	0.20
Cantiani et al. (2019)	Associations with Gamma Power	Not corrected for	At most, 60	0.26 – 0.41	0.25	0.12
Troller-Renfree et al. (2020)	Associations with Relative Theta, Alpha, and Gamma Power	Not corrected for	79	ns	0.22	0.14
Maguire & Schneider (2019)	Associations with Absolute Alpha and Theta Power	Monte Carlo cluster-corrected permutation analysis	At most, 90	0.32	0.21	0.16

* This is the sample size for their correlations, which is sometimes not the maximum size of their samples. ** Significant correlations, in terms of *r*, and for Harmony et al. (1990), canonical *r*. *** This is the smallest significant correlation at their maximum sample size. † This is the power to detect an effect of $r = 0.1$ with their maximum sample size, 80% power, no imbalance, and a *p*-value cutoff of 0.05. ‡ The correlation was reported as two different values, 0.796 and 0.77, but based on the reported *p*-value of 0.002, it could not be 0.796. ‹ The critical *r* for this study was computed at their alpha value. *N*'s are chosen for maximum generosity to the studies, which exaggerates their power in every case.

Table 2. Study Descriptives for Cited SES-EEG Studies[20]

Study	Outcome	Multiple Comparisons	Sample Size	Effect Sizes *	Critical *r*/F **	Power †
Otero (1994)	Associations with Absolute and Relative Delta, Alpha, and Beta Power	Not corrected for	50	6.02 – 19.82	F = 4.04	0.11
Otero et al. (2003)	Associations with Absolute and Relative Delta, Theta, and Alpha Power	Bonferroni without mention of number of tests	42	? ‡	F = 4.08	0.10
McLaughlin et al. (2010)	Associations with Theta, Alpha, and Beta Power	Not corrected for	135	4.61 – 9.42	F = 2.67	0.14
Tomalski et al. (2013)	Associations with Gamma Power	Not corrected for in main analyses	45	3.23 – 4.75	F = 4.07	0.10
Pierce et al. (2019)	Associations with Delta, Theta, Beta, and Gamma Power	Not corrected for	At most, 70	Unclear because of overadjustment, but mostly nonsignificant.	*r* = 0.24	0.13
Cantini et al. (2019) ‹	Associations with Gamma Power	Not corrected for	At most, 81 for *r* and 78 for F	0.23 – 0.30; 6.67	*r* = 0.22; F = 3.96	0.14/0.15
Debnath et al. (2019) ⌂	Associated with Absolute and Relative Theta, Alpha, and Beta Power	Not corrected for in main analyses	45; 138	-0.329 – 0.303; 5.051 – 5.950	*r* = 0.29; F = 3.06	0.10/0.16
Troller-Renfree et al. (2020)	Associations with Absolute and Relative Theta, Alpha, Beta, and Gamma Power	Not corrected for	79	Ns #	0.22	0.14
Brito et al. (2020)	Associations with Beta and Gamma Power	Vague: "FDR correction"	At most, 60	Ns	0.25	0.12
Jensen et al. (2021)	Associations with Theta, Alpha, Beta, and Gamma Power	Not corrected for	At most, 187	-0.204 – -0.176	0.14	0.28

* Significant correlations, or a significant F-value (not to be confused with *f*). Some studies did many different types of tests, so their main ones are highlighted. ** This is the smallest significant correlation or lowest F-value at their maximum sample size. † This is the power to detect an effect of *r* = 0.1/*f* = 0.1 with their maximum sample size, 80% power, and a *p*-value cutoff of 0.05. ‡ Many of their values were left unreported. ‹ These authors also reported a saturated mediation model and, for indiscernible reasons, provided its fit measures. ⌂ Marshall et al. 2008 and Vanderwert et al. 2010 did substantially the same analyses with the same sample and were therefore not analyzed separately. # This was confusing. All relative power relationships with parental education and income were nonsignificant, and the same was true for all but absolute theta's relationship with parental education, which had impossible values. A beta of 0.033 cannot be significant with an SE of 0.166, but 0.33 could, although marginally, in the 'wrong' direction, and not after correction for multiple comparisons. Harmony et al. (1988) could have had usable data, but they did not provide summary statistics for their samples, so SES-ability relationships could not be investigated.

With sample sizes ranging from 13 to 129, the sizes of the effects in this literature were almost certainly *greatly* exaggerated. If we take the minimum significant *r* as the mean effect and assume the real effect is like the relationship between nerve conduction velocity and intelligence,[21] at $r = 0.10$, the best-powered study of the bunch (Brito et al., 2019) still presented an effect that was 1.7 times larger than it should have been, but that study *did not* find significant relationships between any EEG parameters and measured cognitive ability, only with a parent-reported measure of the child's social and emotional development. The next best-powered study in that list would have exaggerated the size of the correlation by 2.1 times. If we calculate effect size exaggeration from power rather than the critical *r*, the results are even less favorable.

As stated in their supplement, Troller-Renfree et al. was sufficiently powered to detect a treatment effect of 0.21 SDs if they had retained 80% of their initial 1,000 participants. With sufficient power to detect an effect of 0.21 at a sample size of 800,[22] they ended up with a total sample size of 435.

Inconsistent Foundations

Troller-Renfree et al.'s study claimed that prior work generally supported positive associations between power at higher frequencies and better cognitive outcomes, whereas power at lower frequencies was associated with the reverse. Yet these studies – and another they cited but did not discuss – failed to follow that pattern. The results from Benasich et al. (2008) were consistent in showing positive associations between absolute gamma power and their various cognitive measures. Likewise, Harmony et al.'s (1990) results went in the same direction, and so did Gou, Choudhury and Benasich's (2011), Williams et al.'s (2012), and Brito et al.'s (2016), with inconsistent significance. But Brito et al.'s (2019) and Maguire and Schneider's (2019) studies were not supportive. The former offered nonsignificant results for their cognitive ability measures despite having the largest sample of the group, and the latter only had one significant cognitive ability-EEG power association: a positive one between low-frequency theta power and working memory. Since theta power is a low-frequency band that the authors argued to be deleterious at higher levels, this finding is striking.

Looking at *all* the studies they cited, an important and undiscussed one was Begus and Bonawitz's (2020) paper *The rhythm of learning: Theta oscillations as an index of active learning in infancy*.[23] This paper came with a plenitude of citations relating theta power to cognitive test outcomes. Several of these suggested greater theta power during memory-related tasks, and that the degree of increase in power and its level are related to greater cognitive performance. Some – in the citations of their citations – even suggested alpha activity during tasks was negatively related to performance, and yet more, that theta activity increased while alpha fell when tasks were difficult.

While these outcomes involved mostly temporal and not resting-state theta or alpha relationships with cognitive performance, they are a cause for concern about how theta and alpha were characterized. The reason for this is simple: many modern theories of intelligence emphasize a potential crucial role of working memory, and even resting-state theta is sometimes positively related with it.[24] Since we do not have temporal EEG information for Troller-Renfree et al.'s study, this concern cannot yet be addressed.

Unfortunately, all these studies suffered from the same kinds of sample size issues as those Troller-Renfree et al. cited positively. As a result, they were similarly uninformative. In fact, all the studies Troller-Renfree et al. cited alongside those mentioned by Begus and Bonawitz suffered from another major, unaddressed issue of potentially being reverse-causal and reflecting behavior rather than causing it or being related due to common causes. Mendelian

Randomization or similar methods could address this issue, but the relevant data does not currently exist.

Troller-Renfree et al. needed a systematic meta-analysis of each of the types of EEG-cognitive ability associations they examined to support the patterns they wanted to observe being beneficial. Moreover, they needed a systematic meta-analysis to assess magnitudes, because if magnitudes of associations are too small, there is no necessary transitivity between EEG enhancements and improved cognitive function. Transitivity is the entire basis of the potential importance of their study; that is, their results might have importance because the changes in EEG power they claimed they observed caused enhanced cognitive ability.[25] If there was no transfer from EEG changes to cognitive ability, then the results were purely of theoretical interest rather than social relevance since no one is concerned with EEG activity on its own.

Confounded Associations

All these studies were uninterpretable for Troller-Renfree et al.'s purposes for another reason: they dealt with cross-sectional associations. In the real world, associations occur for many reasons, and when it comes to neural variables, they are often related to cognitive performance because of genes shared between the neural phenotype and performance. In a study that has been replicated numerous times, Posthuma et al. (2002) found that the positive relationship between general intelligence and brain white and gray matter volumes was accounted for by shared genes, rather than environments.[26] Similarly, the relationship between socioeconomic status and cognitive ability has been found to have a substantial genetic component, and the link between them is substantially severed when genetics are accounted for.[27] Even EEG power is heritable, with most relationships among EEG power at different frequencies attributable to shared genes.[28]

As a result of substantial genetic influence and other potential forms of confounding, how are we to generalize either from experimental effects on EEG parameters to cross-sectional samples or vice versa? Troller-Renfree et al. were at least somewhat aware of these facts, as evidenced by their citation of Wax (2017), who argued that "the so-called neuroscience of deprivation has no unique practical payoff… Because this research does not, and generally cannot, distinguish between innate versus environmental causes of brain characteristics, it cannot predict whether neurological and behavioral deficits can be addressed by reducing social deprivation."[29]

Unrepresentative Samples

Though genetic confounding loomed large, it was not the only sampling issue. These studies generally sampled populations that were in some way unusual or limited with respect to the full ranges of cognitive ability and socioeconomic status. For example, Benasich et al. (2008) used a sample within which 85% of parents were in the highest socioeconomic strata and the other 15% were middle-class, and nearly a third of the sample came from families with histories of language impairment. In Maguire and Schneider (2019), the low-income sample had an IQ (99.2) that was indistinguishable from the population average (100). The sample used by Gou, Choudhury and Benasich (2011) had an average IQ of 113 – a standard deviation above the general population! Williams et al.'s (2012) sample may have had lower than average cognitive development because of the direct effects of the participants' congenital heart disease, or because of genes shared between low cognitive ability and heart disease. In the UK Biobank dataset, these phenotypes are related such that angina, chronic ischemic heart disease, major coronary heart disease events, and a broadly defined ischemic heart disease are all significantly negatively

genetically correlated with fluid intelligence scores. Other reviews have also highlighted frequent cognitive deficits among individuals with congenital heart disease, while noting that phenomena such as enhanced mosaicism in trisomy 21 and greater numbers of copy-number variants, commonly found in congenital heart disease sufferers, were related to worse cognitive outcomes.[30] Some studies even involved individuals with IQs below 90 – only two-thirds of a standard deviation from the mean – being removed from the samples used in their analyses. Given that the studies cited by Troller-Renfree et al. had problematic and unusual sampling, we have little reason to believe they serve as a solid evidentiary basis for theorizing about or predicting population outcomes.

Multiple Comparisons

Finally, only three of the nine studies corrected for multiple comparisons, and of the three that did correct for them, only one (Williams et al. 2012) did so in a standard and verifiable way. The problem of multiple comparisons is important, and amenable to graphical explanation. Figure 1 illustrates the relationship between the number of statistical tests and the probability that at least one of the results is significant with a significance threshold of 0.05 *if there is no effect*.

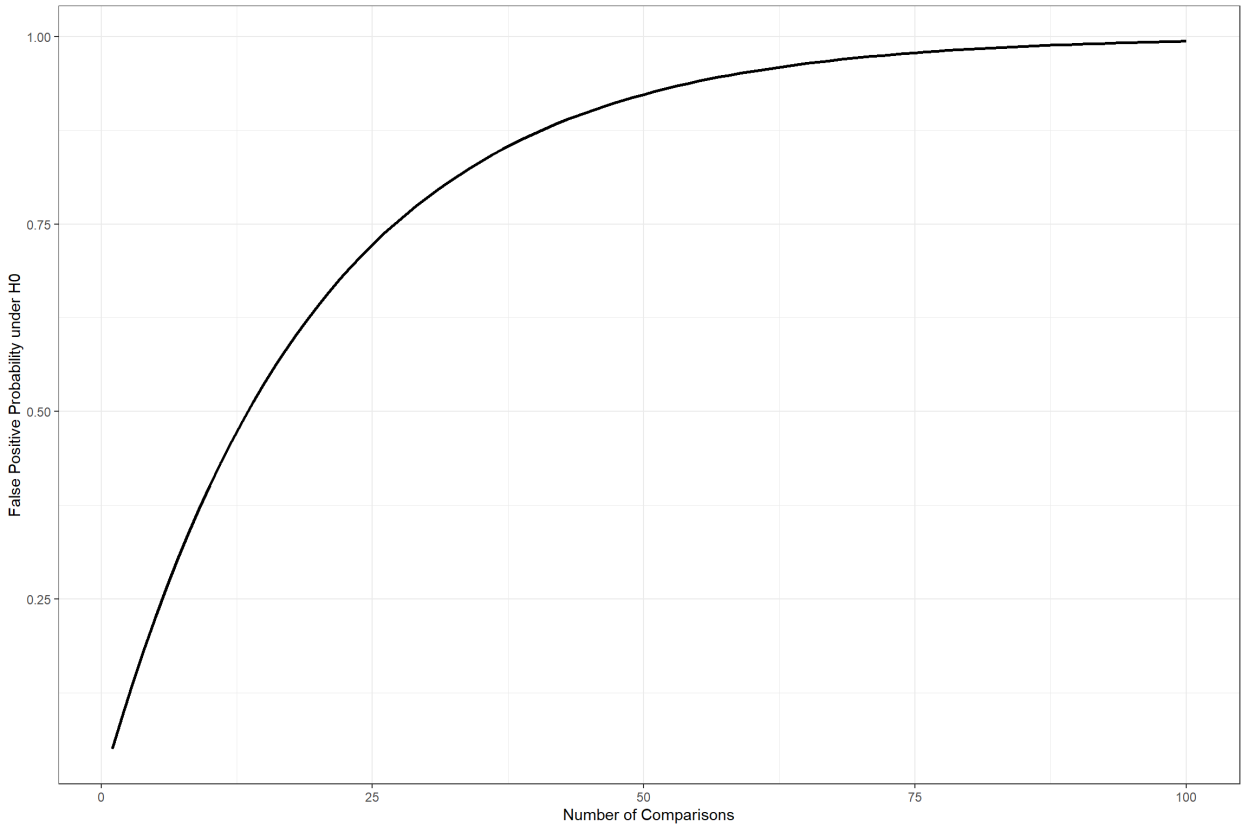

Figure 1

The issue with small sample sizes is also open to graphical explanation. Figure 2 illustrates the relationship between sample sizes and the average sample correlation conditioned on reaching a 5% significance threshold (the mean significant correlation) in simulated data with different true underlying correlations. The mean sample sizes in Tables 1 and 2 were 89 and 67, respectively.

If the true correlation is typical of the correlations observed in large, reliable studies in neuroscience (closer to 0.10), correlations in those studies were exaggerated by 2.6 to 3 times, assuming balanced samples, a lack of measurement error, and no heteroskedasticity, high-leverage outliers, or nonnormality, and therefore overestimates power while underestimating effect size exaggeration.

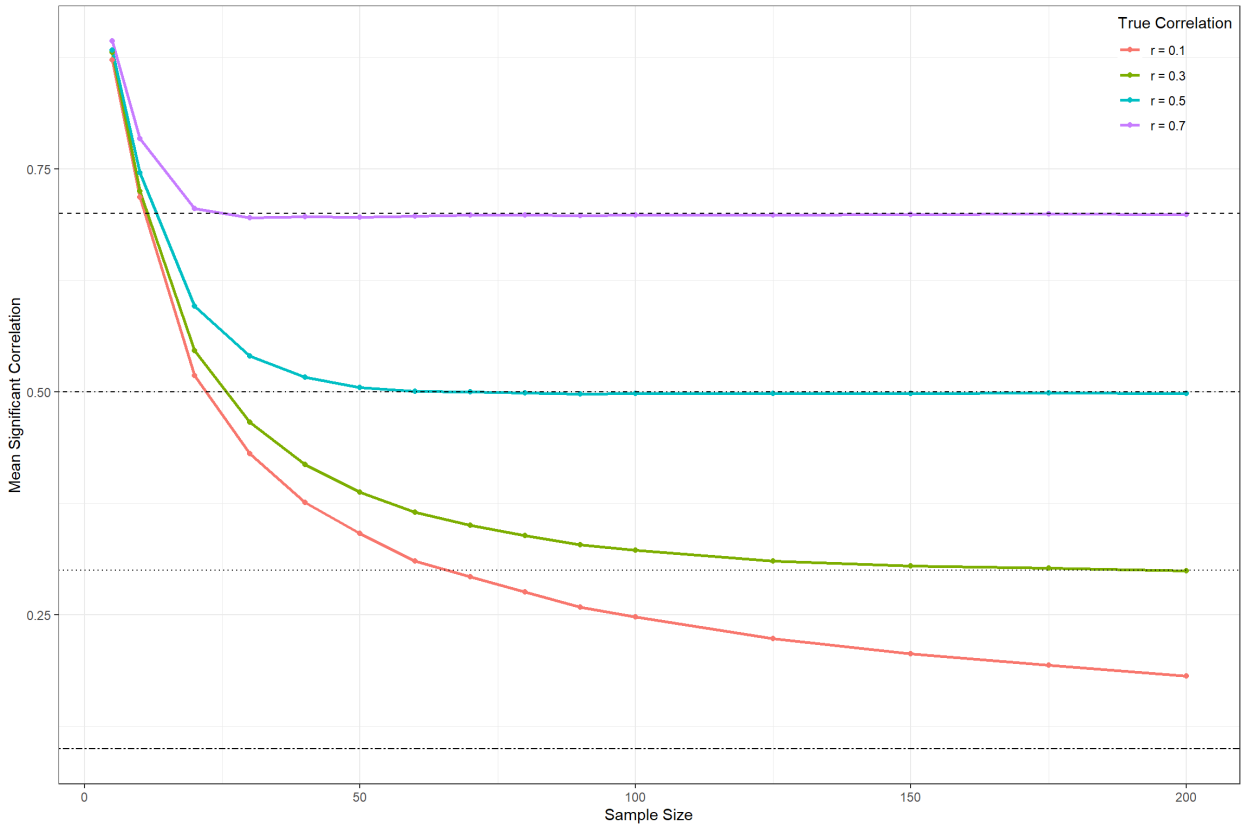

Figure 2

If we perform five tests with a *p*-value cutoff of 0.05 and there are really no differences, the probability that at least one *p*-value is below 0.05 is $1 - (1 - 0.05)^5 = 23\%$. With ten tests and no correction for multiple comparisons, the likelihood of a false-positive is 40%. The interpretation of a *p*-value as the probability of a more extreme result given the truth of the null relationship drifts as more tests are conducted. All the studies without corrections for multiple comparisons in Table 1 had more than ten tests, so the results should be considered to have been run with unacceptably liberal *p*-value cutoffs.

The only study that can realistically be generalized from to support the patterns Troller-Renfree et al. wanted was by Marshall et al. (2008). That study featured comparisons of absolute and relative alpha and relative beta power between groups of Romanian children who were randomized between institutional and foster care settings. The *p*-values were Bonferroni-corrected, the sample included 41 institutionalized children and 49 who were placed into foster care, and the authors observed no significant differences in power measures between a group of 22 placed into foster care before two years old, a group of 27 placed after two years of age, and the 41 institutionalized children. Placement age likewise was not significantly related to absolute theta, alpha, or beta, nor to relative theta or beta, though it was significantly negatively related to relative alpha power. None of the mean differences in absolute or relative power between the

institutionalized or foster care groups were significant. It is hard to infer anything regarding EEG patterns of socioeconomically meaningful interventions from this study.

The studies that Troller-Renfree et al. proffered to support a strong interpretation of their results were weak. They utilized small samples, had inconsistent findings, and were sampled in ways that make generalization difficult. Even if they had found significant effects of their intervention on EEG power parameters, there was little to no reason to think these would elicit a causal impact on behavior.

Problems With the Study Itself

Once these methodological issues are considered, it becomes clear that Troller-Renfree et al. misled their readers about the ironclad nature of their results. In fact, their intervention yielded no reliable statistically significant effects. They observed no significant differences in group means of the absolute or relative power parameters and, after including many covariates, they arrived at treatment effects ranging in absolute value from 0.02 to 0.26. Prior to correction for multiple comparisons, the only significant effects were found for absolute beta and gamma power. With *p*-values of 0.02 and 0.04, respectively, the effects were at best marginal, and as might be expected given a marginal result among many tests, they stopped being significant when the authors corrected for multiple comparisons.

With no significant results, the authors still positively interpreted their study. They claimed that their "findings underscore the importance of shifting the conversation to focus more attention on whether or how income transfer policies promote children's development," but this claim was not supported by their results.[31] Moreover, these findings probably ought to be given less prominence because they did not stick to a preregistration plan in the way claimed by the authors.[32]

The Unregistered Tests May Have Contradicted the Authors

The largest effect Troller-Renfree et al. observed – for absolute beta – was *not* preregistered. They attempted to justify assessing effects on beta in their supplement by saying that studies at the time of preregistration did not link income to EEG power, but studies since then had indicated that income *was* in fact linked to beta activity. They cited two studies in support of that view.

The first study, Brito et al. (2020), featured a sample of 94 infants and reported that maternal education and the income-to-needs ratio were significantly associated with EEG beta power (*p*'s = 0.025 and 0.048); however, neither survived the authors' correction for multiple comparisons. Furthermore, in their regressions with adult word count, conversational turns, and child vocalizations, there was one significant relationship in 15 regressions, and these same regressions included no significant relationships between the income-to-needs ratio and beta in any region or whole-brain gamma power. Infant auditory comprehension and expressive communication were significantly related to one another, but not to either measure of income or maternal education. Other relationships with those measures were unclear because the authors' supplement was improperly uploaded and is currently just a .zip file containing a .xml file that includes nothing.[33]

The second study, Jensen et al. (2021), took 210 6-month-old children from the "Cryptosporidium Burden Study," and another 210 36-month-old children from the "Performance of Rotavirus and Oral Polio Vaccines in Developing Countries" study. After attrition, their sample sizes totaled 160 and 187, respectively. Among the 6-month-olds, wealth, maternal education, family care, and maternal stress were not significantly related to any EEG

parameters; among 36-month-olds, maternal education and family care were unrelated to EEG parameters, while maternal stress was significantly positively related to frontal and central theta, and wealth was significantly *negatively* related to all beta and gamma power parameters. In their regressions, there were no significant absolute or relative EEG relationships with maternal education, wealth, family care, or maternal stress in the 6-month-old group. In the 36-month-old group, the regressions yielded the same results for relative power, but there were significant positive relationships between absolute theta and maternal stress, and significant negative relationships between wealth and absolute beta and gamma power.

The inconsistencies of both results with those of Troller-Renfree et al. are obvious: one study produced nulls and may have done the same with cognitive measures, and the other produced effects in the opposite of the direction of Troller-Renfree et al.'s intervention effects, alongside a bundle of null associations. These hardly serve to justify their unregistered analyses, and thanks to the specific results in these citations and their generally larger size than the other studies Troller-Renfree et al. cited, they further reinforce the inconsistency of this literature with their results and the potential incommensurability of experimental and cross-sectional EEG associations.

The Study was Reliant on Typically Unimportant Analytic Decisions

Two additional unregistered methodological choices were critical to Troller-Renfree et al.'s results. The first revolved around the fact that they produced their estimates using robust regressions. However, a glance at their code revealed that the regressions were not quite robust. They used what are known as HC1 standard errors – the default in the program they used to analyze their data, Stata. These standard errors fare poorly under scenarios where certain participants exert high leverage on estimates, and with even moderate sample sizes like theirs, the false-positive rate of HC1 robust regressions is considerably above the nominal rate of 5%. Rerunning their analyses using HC5 standard errors – which perform better with small samples, high leverage, nonnormality, and severe heteroskedasticity – I found no significant results even before accounting for multiple comparisons.

The second methodological decision involved how they computed their effect sizes. In evaluating RCTs, it is common to divide the effect by the standard deviation (SD) from the control group. It is also common to divide the effect by the pooled SDs from either group, weighted or unweighted by the sizes of the samples. The authors did not offer any principled justification for one effect size calculation method over any other, but they did choose the only one that would deliver them significant results. In choosing to divide by the SD from the control group, they bought themselves significant *p*-values for the absolute beta and gamma power effects, where any other method would have rendered their effects nonsignificant. In the case of absolute alpha power, had they used either SD pooling method, the effect would have been negligibly larger and still nonsignificant.

Though not a methodological decision, the authors also included an interpretive one that was biased towards a positive interpretation of their results. The authors were inconsistent about the interpretation of results that were marginal. Twice in the main study, they noted that effects were "at the margins of statistical significance," and they did the same three times in their supplement. So-called "marginally significant results" are often those that are *just nearly* above the *p*-value cutoff. Researchers will frequently interpret these, but they seem to nearly universally neglect that the "marginal" label must be two-sided: if a result is marginal at a *p*-value *above* the researcher's chosen cutoff, it is equally marginal at the same distance *below* that cutoff. These authors called *p*-values as high as 0.10 marginal; if we were to give those findings

a significant interpretation, they ought to consider at least anything greater than 0.01 to also be marginal. This turns every significance test they ran at best marginal, and more realistically, nonsignificant. They cannot have it both ways: effects cannot be marginally significant if they are not, on the other end, also marginally nonsignificant.

These decisions are normally unimportant, and they possibly represent the study's authors using defaults without giving them any thought. For a study that produced borderline results, these decisions were critical. For a well-powered RCT, this sort of methodological decision would not usually affect results, but when they *determine* the result, the study must be considered suspect.

In their supplement, Troller-Renfree et al. produced a series of region-specific EEG effects. Six of the thirty-two region-specific associations were marginally significant (*p*-values between 0.01 and 0.04). These associations did not survive correction for multiple comparisons when I did them correctly (i.e., across bands), but three of them narrowly remained under 0.05 when they were done incorrectly (i.e., only within frequency bands, unlike the analyses in the body of their paper). The reason this result is notable is because the study's senior author, Kimberly Noble, claimed that this *post hoc* vindicated their results in an update to the *Vox* story on this study.[34] She "argues that the results are robust because of what is known as 'regional analysis' – analysis that looks at where in the brain differences between the high-cash and low-cash children showed up" and is quoted as saying that "if they had come from, say, parts of the brain that support vision, we might have been very skeptical that we were seeing something meaningful. But they came from parts of the brain that are critical for supporting higher-order thinking."

There are four major problems with this kind of thinking. First, according to their studies and reasoning for assessing beta associations (noted above), the effects they observed went in the wrong directions for the beta power band, and potentially also for the gamma band. Second, there are no validity studies that support their proposed region-specific effects with a reasonable degree of certainty. Third, this multiplies the number of tests, and we ought to take it for granted that, were they all significant together, they would have put them up front and center, since they now seem to be doing something similar with more restricted analyses in order to protect their study from criticism. And finally, the reasoning is plainly faulty: claimed region-specific associations with brain variables ought to almost never be trusted. The reason this is the case is simple and has to do with sample sizes and power. When you have low power, your effect sizes must be exaggerated. When you have low power and you find a region-specific association, it is almost certainly because you had one region with an extreme and unrealistic result, while the others were more realistic, and the study was not powered to detect if those differences existed at all. Many region-specific associations with fMRI have been like this: they are the results of low-power studies being exploited to capture chance phenomena.[35]

Other EEG Analyses

The authors went on to find that there was a significant treatment effect ($p = 0.02$) on summed high-frequency bands. But they likely didn't expect this to happen per their own cited beta (and maybe gamma) validity studies. This test was not corrected for the many comparisons it was featured among and all it could do was recapitulate group differences in the high frequency bands, with additional power to reject a lack of difference between groups because summing them reduces measurement error, while pulling many assumptions into the analysis and removing analytic clarity.

In another analysis, they performed their focal analyses across all power bands but with log-transformed absolute and relative powers, because some people in psychophysiology had argued for logging variables prior to analysis. This rendered all their estimates nonsignificant prior to multiple comparison correction. Because of how the authors are now backtracking, it is reasonable to suspect they would have both justified and presented log-transformed results if they were more significant than their untransformed results.

Researcher Degrees of Freedom

The idea that researchers would only present their significant results and methods that supported their desired interpretations while moving the rest of their results to their supplement or out of the paper entirely is sometimes dubbed "researcher degrees of freedom." Preregistration reduces the degree to which authors can positively misreport or chop up their findings, but there is usually still leeway to selectively report and highlight results. The proposition that splitting data in different ways would validate nonsignificant findings because some number of the subset effects are significant is a spurious one. The result of doing this is more comparisons and an effectively lower significance threshold, which means more exploitation of chance.

This search for significant results in subsets of data has led to the popular repetition of the phrase "the difference between 'significant' and 'not significant' is not itself statistically significant."[36] This phenomenon appears in situations such as when an effect is significant in one group but not another and authors interpret that observation as evidence of a difference in the size of the effect by group. But the difference in the effects may not be significant; the study may simply be underpowered for the effect to be significant in one group while the other may have had a more extreme result than they should have by chance alone.[37] The proper way to test this difference is to test for an interaction in a singular model or to directly compute the significance of the difference in effects. One reason this sort of thing is not done more often may be because it requires *much* larger sample sizes.[38]

Many papers have dealt with this kind of data abuse to wring out significant results. For example, Miller and Sanchez-Craig (1996) gave tongue-in-cheek advice to researchers to do things like "use liberal definitions of success" and "always declare victory regardless of findings."[39] In more recent times, Simmons, Nelson and Simonsohn (2011) have noted how "flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not."[40]

Wicherts et al.'s (2016) checklist of researchers' degrees of freedom has considerable relevance to Troller-Renfree et al.'s paper and their attempt to salvage nonsignificant main results with reference to exploratory secondary ones.[41] Examples they listed include:

- Studying a vague hypothesis that fails to specify the direction of the effect
- Creating multiple manipulated independent variables and conditions
- Measuring the same dependent variable in several alternative ways
- Failing to conduct a well-founded power analysis
- Measuring additional constructs that could potentially act as primary outcomes
- Selecting the dependent variable out of several alternative measures of the same construct
- Trying out different ways to score the chosen primary dependent variable
- Operationalizing manipulated independent variables in different ways (e.g., by discarding or combining levels of factors)

- Choosing between different statistical models
- Choosing the estimation method, software package, and computation of standard errors
- Failing to report so-called "failed studies" that were originally deemed relevant to the research question
- Misreporting results and *p*-values
- Presenting exploratory results as confirmatory (Hypothesizing After Results are Known, HARKing)

All of these can be fine in specific scenarios, but they are also pathways towards allowing the misuse of statistics and the misrepresentation of results. The *post hoc* presentation of region-specific analyses with dubious expectations for those results is one such abuse example.

Did the Intervention Affect Cognitive Development?

Troller-Renfree et al. also tested the effects of cash transfers on a questionnaire about language milestones and its relationships with EEG parameters. Their language milestone assessment, the Ages and Stages Questionnaire 3 (ASQ-3) is a maternal report instrument used to assess linguistic development in young children.[42] The authors noted that this instrument had been strongly correlated with another development screener, the Battelle Developmental Inventory, and that it had also been found adequately reliable.

In the subset of the study with EEG data, there was a significant treatment effect on the ASQ-3 ($p = 0.03$), but it was marginal, per the authors' implicit definition. If coupled with other tests, it could not survive correction for multiple comparisons. The authors went on to test the treatment effect again, but this time using the full sample of 900. When they did this, the effect shrank enough to become nonsignificant ($p = 0.21$). The difference between these coefficients was not significant, but the authors did not test to see if that was the case.

The authors attempted to explain their nonsignificant effect in a few ways and did note that some of the explanations were not especially plausible. First, they explained the finding as the result of a possible difference response in patterns between in-person and over-the-phone responders. They rendered this less likely by noting that ASQ-3 scores did not differ by collection method ($p = 0.67$). Second, they suggested differences between the control and treatment samples with respect to levels of language input could have given rise to a reduced effect because the pandemic equalized language input due to increased time at home for everyone regardless of group. And third, systematic differences between the EEG and total samples that interacted with the cash gift may have negatively affected the total sample relationship. They reduced the probability of this explanation by noting that their non-response weight analysis better matched the EEG sample to the total one and the same pattern of EEG effects from before the weights were applied was still observed.

These explanations did not apply to the same things despite being listed together. The first and third explanations had to do with the difference between the EEG and total samples, the second, with the existence of the effect. This is relevant because it is peculiar, and more importantly, because all these explanations are, in a sense, immediately testable by the study's authors. But even though they released their data with the study, *we* cannot test it.

The way to test all these explanations for the differences is to assess something known as measurement invariance. Measurement invariance is the condition where a questionnaire or assessment is unbiased when used in different groups.[43] Without bias, differences between groups reflect the construct the instrument is supposed to measure.

If all the utilized ASQ-3 items were provided, we could assess the first and third explanations by assessing measurement invariance between the EEG and total samples. If they were correct, measurement invariance would be violated. The second explanation is testable by assessing measurement invariance between treatment and control groups – if it (or another explanation) was true invariance would be violated.

The method used to test the second explanation must be applied *in general*. To assess intervention effects on psychological outcomes and interpret the effects as alterations of a particular psychological construct requires invariance between the treatment and control groups. To interpret changes *over time* in terms of a given construct target requires longitudinal invariance, and to interpret differential changes in a construct, invariance over time needs to be assessed across groups.

If the effects of treatment are heterogeneous, this will cause a violation of the strict phase of testing invariance, in which the equality of *causes* of differences aside from the construct are tested. This would yield an important note: if different influences existed between groups, the group with the greater variances would have additional influences on responding. Failing to support this level of invariance might make some of the intervention effect meaningless.

Consider intelligence. An RCT with an effect on all the scores from a test, in which strict invariance was violated, means something *besides* intelligence was affected, either alone or in addition to effects on intelligence. In intelligence research, it is a stylized fact that most – and often virtually all – of the predictiveness of IQ tests is due to the general factor of intelligence, *g*. If your intervention affects *g* and question-specific performance, your intervention will affect things *g* leads to, like higher rates of learning on the job or a greater capacity to solve a variety of complex puzzles. When it affects performance specific to a certain test, this is not as likely to affect external outcomes or to yield success in the future. If you get very good at addition, that by no means makes you better at public speaking; if you improve your intelligence, that should. If I give a sample all the answers to an IQ test and I make the control group take the test normally, the treatment (giving out answers) does not affect intelligence, it only affects the scores, and that does not mean the kids with the answer key became more intelligent. Psychological RCTs *always* need to be measured with this in mind: measurement matters, and if it is not done, the results are ambiguous and changes in measured outcomes might be useless.[44]

Finally, the authors threatened the interpretation of their study by correlating ASQ-3 scores with the whole gamut of absolute and relative EEG power estimates. The *p*-values ranged from 0.39 to 0.95: they were all *very* nonsignificant, despite being much larger than the studies from Table 1. They offered more measurement-based explanations and the possibility that the links would grow with development. But if this were the case, and they were asking us to expect totally different effects on cognitive outcomes at older ages, why not also have totally different patterns and magnitudes of EEG effects?

Addressing Attrition and Missingness

Troller-Renfree et al. also tested what inverse probability weighting (a technique to allow the computation of effects when the samples are more equal in terms of baseline traits) and non-response weights (intended to address missingness issues) did to their estimates. These weighting schemes made their results universally nonsignificant, so they must have known before presenting their results that it was wrong to present any of them as significant when they were not robust to reasonable and informative analytic choices like these.

Did the Study Show Parents Spent the Money Well?

Per the *New York Times'* coverage, economist Lisa Gennetian – one of the study's co-authors – said "the results indicated the parents could be trusted to make good decisions" with free money. The study as presented did not and could not show that money given to parents was spent well. The lack of effects means we cannot tell if giving people money has improved anything for their kids. More importantly, the controls likely removed from the analysis much of the tendency of parents to misbehave. Among the long list of controls were maternal mental health, maternal smoking and alcohol consumption during pregnancy, and a variety of other life history speed and misbehavior correlates. When parental misbehavior is treated as a construct that can be measured by things that indicate it, it is apparent that the authors overcontrolled by controlling for at least part – and probably a large part – of the variance in parental misbehavior. At the very least, they overcontrolled if the goal was to show money was spent well. Perhaps parental misbehavior was an important source of heterogeneity! Because the effects were so nonsignificant, the present data cannot be used to test that suggestion.

Should We Expect Effects to Grow?

Vox quoted psychology professor Allyson Mackey on Troller-Renfree et al., "My prediction is that the brain effects of cash transfer will grow as kids grow up."[45] If this idea is relegated to effects like those on EEG parameters, then, as we have seen, it may well be useless. If, however, it also includes downstream behavioral effects from those brain ones, then we have an existing evidence base that contradicts it.

A nigh-universal phenomenon known as the fadeout effect plagues interventions designed to improve people's educational (and other) outcomes. In a 2015 review, John Protzko showed that interventions as diverse as intensive preschool programs and vitamin A supplementation, which initially appeared to elicit effects, all faded in their effects at subsequent assessments.[46] Bailey et al. (2020) also conducted an excellent and accessible review of this phenomenon.[47]

In addition to the fadeout effect, there is a phenomenon in intelligence research[48] known as the Wilson Effect, by which systematic environmental effects fade and genetic effects increase with age.[49] These alone are enough to throw the idea that the intervention effects should grow into disarray. If people are arguing that continued exposure to higher income will accumulate larger effects, then they have a gigantic burden of evidence, and, hopefully, they will assess fadeout after the intervention inevitably ends.

The New York Times quoted Representative Suzan DelBene (D-Wa.) as saying that Troller-Renfree et al. showed that "investing in our children has incredible long-term benefits."[50] The evidence for this claim is nonexistent. The study covered one year, and there was no reason to think effects would persist or generalize such that they would become meaningful by modifying participants' actual behavior and cognitive abilities. Long-term benefits can never be taken for granted.

It's a "Contribution"

PNAS has two submission categories: the typical one and "Contributions."[51] Contributions are available as an option for National Academy of Science members, and they have been controversial since they have been alleged to make peer review easier.[52] This kind of publication involves the article submitter selecting their own reviewers and gathering up their own reviews. Troller-Renfree et al.'s article was submitted by Greg Duncan, and he selected Martha Farah and Joan Luby to review.

These reviewers had prior publications with some of the coauthors of Troller-Renfree et al. Martha Farah has published at least six articles[53] with Kimberly Noble, and Joan Luby[54] published an article[55] with Troller-Renfree and Nathan Fox. Normally, this would not be concerning. Fields are sometimes small and subject matter experts are often few, but it is an issue when you get to choose your reviewers and they are personal friends and colleagues. The freedom of the review process from bias in this situation is hard to maintain.

Most Interventions Have Small to Nil Effects

To increase the believability of their large and probably exaggerated effects, Troller-Renfree et al. stated that "the observed effects in the alpha-, beta-, and gamma-bands are similar in magnitude to those reported in other large-scale environmental interventions. For example, a meta-analysis of 747 randomized control trials of educational interventions targeting standardized achievement outcomes found an average size of 0.16 SDs." This was a great overstatement. The study (Kraft, 2019) was interpreted as charitably as possible to their own views.[56] The effect of 0.16 the authors noted was nonsignificant and was likely to result from small study biases like those mentioned above. In fact, small studies *did* have much larger effects than larger studies in this meta-analysis, and there were fewer studies the larger the samples got. This was also consistent with the observation that the median effect sizes were lower than the mean ones and their distance decreased with increasing sample size. The meta-analytic effect size weighted by study sample sizes was 0.04, and not significantly different from zero. Because small study effects still exist even when studies are weighted, this ultimate effect is still probably upwardly biased. When looking at math and reading separately, the same phenomenon occurred, and the same result happened. These educational interventions were not what they were cracked up to be.

Educational Interventions

Kraft's 2019 study was not the only one on educational interventions. The 2010 Head Start Impact Study conducted by the Office of Planning, Research and Evaluation looked at effects elicited by the nationally-implemented, well-funded, and widely celebrated Head Start program for three- and four-year-old children.[57] Using a *p*-value cutoff of 0.10, they found significant positive effects on eight of fourteen cognitive tests given in the year of the program for the three-year-old children, and this reduced to two of fourteen by the next year, at age 4. In kindergarten there were nineteen tests administered, with one significant positive effect, and one significant negative one, while, in the 1st grade, only one effect of twenty-two remained significant. Because this was not significantly affected during Head Start and its effect only became significant amongst a pile of nonsignificant effects, this was probably noise. For the four-year-old group, seven of fourteen scores were significantly positively affected in the Head Start year, but effects disappeared in kindergarten, and one effect became significant again in first grade. In a particularly informative analysis by Lortie-Forgues and Inglis (2019), they found that the typical educational intervention had a minute effect (mean = 0.06 SDs) and was unconvincing in Bayesian terms.[58] Though their analysis included fewer studies than Kraft's, their analysis was probably better because it mitigated publication bias by only using studies funded by the Education Endowment Foundation and the National Center for Educational Evaluation and Regional Assistance, both of which require all trials they fund to be published and to follow standardized reporting practices.[59]

As shown by Duncan and Magnuson (2013), the effect sizes associated with early childcare programs declined over time.[60] Repurposing their data, I found a correlation of r = -

0.37 between a study's precision and its effect size, meaning that the more power the study had, the smaller its effect. Adjustment for publication bias rendered the meta-analytic effect nonsignificantly different from zero, and this was true *even though* the authors recoded outliers via Winsorization. This was qualitatively like the result of a Brookings report from 2018 that found state-level changes in prekindergarten enrollment were not associated with changes in state-level achievement scores.[61]

Cash Transfer Effects

Although Troller-Renfree et al.'s study was the first to assess the effects of unconditional cash transfers on EEG parameters, it was not the first to assess effects on cognitive ability. Baird et al. (2013) found that unconditional cash transfers did not have a significant effect on test scores (0.04, 95% CI: -0.04 – 0.12), and neither did conditional cash transfers (0.08, 0 – 0.16), but together, they yielded a marginally significant effect (0.06, 0.01 – 0.12). On the other hand, both types of cash transfer had considerable effects on enrollment and attendance, but there was substantial evidence of publication bias that the authors did not compute the effects of. Luckily, since the data were available, I was able to find that the results did not survive weighting studies by the square root of each study's sample size. To explain away the nulls, one could argue that these interventions did not offer large enough cash subsidies or that they were too small, or enrollment was range restricted. But to the extent that is true, it just increases our uncertainty about cash transfer effects since they have not been found to have meaningful effects on test scores, much less psychometrically meaningful variables like intelligence, memory, or verbal ability.

The World Bank also evaluated several other cash transfer programs.[62] The studies in their Table 3.1 ranged in timeframes from one to ten years. Three of them were unconditional, while the other five were conditional cash transfers, with sample sizes ranging from 106 communities to 2,069 kids and were all either RCTs or randomized phase-ins. Unlike the unconditional transfer studies, the conditional ones supplemented cash transfers with schooling and health check-ups. In Paxson and Schady (2010), language, fine motor function, and socioemotional skills were not affected, and there was a positive effect on one of the many Woodcock-Johnson subtests, long-term memory (0.18 SDs). In Fernald and Hidrobo (2011), there was a significant effect on the MacArthur Language Test (0.15 SDs) for rural kids, but no effect on urban ones. In Lopez, Boo and Creamer (2018), ASQ-3 scores were improved by 0.13 SDs and communication was improved by 0.18 SDs. Barham, Macours and Maluccio (2013) found 0.15 SDs of cognitive effects and no significant socioemotional effects. Fernald, Gertler and Neufeld (2008) found effects of 0.11, 0.10, and 0.09 on long- and short-term memory and visual integration when the cash transfer was doubled, and they also found a 0.18 SD effect on vocabulary, but no significant effects on fine motor function. Fernald, Gertler and Neufeld (2009) found no significant effects on cognitive ability or language skills, but they did observe a 0.14 SD reduction in behavioral problems. The only study in the batch that involved multiple follow-ups (Macours, Schady and Vakis, 2012), showcased fadeout. At kids' first follow-ups, there were significant effects on short-term memory (0.15 SDs), vocabulary (0.23 SDs), and other language parameters (0.14 SDs), alongside a 0.13 SD socioemotional effect, and no significant effects on fine motor function. At the second follow-up, the short-term memory effect declined to 0.09 SDs, the effect on vocabulary was not measured, the language effect dropped to 0.09 SDs, and the socioemotional effect dropped to 0.01 SDs, while, inconsistently, effects on fine motor function became significant at 0.16 SDs.

The same World Bank report also covered three cash transfer studies in greater depth. These studies were conducted in Niger, Mexico, and Colombia, and the Colombian study was also covered by Baird et al. (2013). Initial sample sizes ranged from 1,420 to 6,856 children. The Colombian study compared three treatment groups to a control group, who received conditional cash transfers. The first treatment group received psychosocial stimulation, the second received micronutrient supplementation for anemia, and the third received both. There were no significant differences in any of the treatment group comparisons with the control group in height, weight, socioemotional development, and gross and fine motor function. There was a significant effect of treatment relative to the cash group for the Bayley-III cognitive test and receptive language at the first follow-up in the comparison of the first treatment and the control group. None of these effects remained at follow-up. The Mexican trial compared two treatment groups to a control group. The control group received money and the treatment groups received money plus available (group 1) or available and encouraged (group 2) parenting group sessions. There were no significant differences between group 1 and the control group, but there were considerable impacts on memory, full-scale IQ, and verbal scores for the second treatment group. There was no follow-up. The Niger study compared a control group given unconditional cash to a treatment group given conditional cash plus parent training about nutrition, psychological stimulation, health topics, and sanitation. There were no significant effects on cognitive development, but children's socioemotional problem index and sociality scores increased. There was considerably less self-reported illness, more handwashing, a greater likelihood of receiving all recommended vaccinations, and iron levels improved.

Nutritional Interventions

Though they usually have nothing to do with cash transfers, it can be informative to look at other large intervention evaluations to see what sorts of things elicit effects and in what magnitudes. In Nepal, Dulal et al. (2018) ran a double-blind RCT where groups of mothers were given prenatal multiple micronutrient supplementation in the treatment group, and iron and folic acid only in the control group.[63] At twelve years of age, there were no effects on the children's cognitive ability. Behrens et al. (2020) reviewed whether vitamin B affected rates of cognitive decline and found no effects, despite vitamin B mechanisms supposedly being well-defined and the effects plausible at the time of writing.[64] Similarly, a large-scale follow-up of two double-blinded, placebo-controlled, cluster-randomized trials of vitamin A effects on prenatal and newborn children by Ali et al. (2017) yielded nonsignificant effects across the board.[65]

Deworming

Deworming charities have been a cause célèbre among effective altruists for some time now. Welch et al. (2019) reported the results of a systematic review of deworming effects on a variety of outcomes including cognitive ability, height, and weight.[66] Despite parasitic worms often being debilitating and the mechanisms through which they should impede development being abundantly clear, there were no significant meta-analytic effects on weight, height, or cognitive ability. Welch et al. compared their results to earlier systematic reviews by Taylor-Robinson et al. (2015) and Welch et al. (2016) that, altogether, found one marginally significant meta-analytic effect on weight and none for height or cognitive ability.[67]

Brain Training

Brain training interventions have ebbed and flowed in popularity in the intelligence testing literature. Methods to improve working memory, such as the Dual N-Back, were at one

time very popular and, based on published results, even seemed to raise intelligence.[68] However, the case of the Dual N-Back is like those of other methods in that, once greater rigor was applied, the effects stopped generalizing or disappeared.[69] A large second-order meta-analysis of brain training studies (Sala et al., 2019) came to similar conclusions.[70]

Will the Effects be Larger with Greater Cash Transfers?

One may suspect that cash transfers larger than those distributed by Troller-Renfree et al. might have a desired effect, thus putting a more positive gloss on the results of their paper. In general, however, interventions cannot simply magnify their effects by increasing the amount of stimulus provided. For example, if we set up a daycare program, it is not as though adding twelve hours of daycare supervision would improve children's cognitive outcomes even if a study showed that one hour was associated with better outcomes. Similarly, if children were deprived of iodine, an intervention that gave children enough to poison them would not yield improvements over providing the recommended daily allowance.

Certain designs can appear to be scaled-up versions of the present interventions. For example, the Moving to Opportunity program that relocates people to high-quality neighborhoods has a staggering economic value. And yet, Ludwig et al. (2013) found that Moving to Opportunity had no effect on a mathematics and reading assessment, attained education, or almost any other outcome.[71]

An even more impactful form of intervention is adoption. The largest adoption studies afford modest effects[72] – on the order of one-third of a standard deviation – despite very large environmental improvements.[73] But effects like this are only a fraction of the apparent relationships with socioeconomic status found within nonadoptive families, in which siblings and parents share genes rather than just environments. To understand this, take the following plot, derived from Odenstad et al. (2008).[74] The plot includes three groups in Sweden: Korean adoptees, non-Korean adoptees, and biological children. These groupings are relevant because, uniquely for adoptions, among Korean adoptees, would-be parents cannot choose the child they receive, so selection effects are removed and, consistent with this, the effect of age at adoption on attained IQ disappeared in the Korean – but not the non-Korean – samples. Despite considerable adoption effects on attainment, the effects were not large enough to generate a systematic relationship between adoptive family socioeconomic status and cognitive ability.

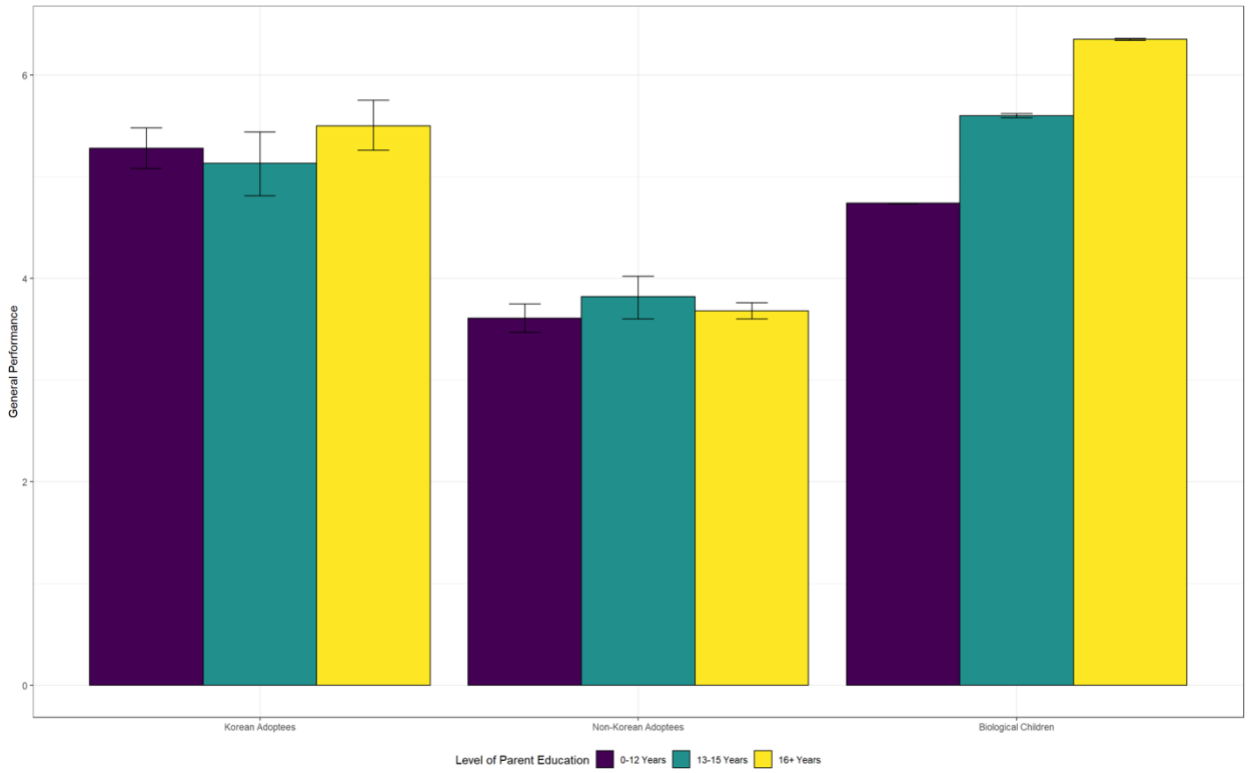

Figure 3

This plot clarifies that adoption effects, as large as they may be, are a fraction of the realized differences related to socioeconomic status in the general population when the link between family socioeconomic status is confounded with genes. Additionally, there is just no reason *presently* to suggest the effects will grow considerably past some likely already near enough point, or that they will *ever* reach the level expected based on cross-sectional outcome-socioeconomic status associations. Extensive intervention fadeout at all ages might also signal a long-term lack of efficacy after the intervention. In any case, Troller-Renfree et al. and commentators in this area should not generalize from biologically, or otherwise considerably, confounded associations.

A potential counterargument is that effects earlier in life will better persist and have larger effects than ones based on later interventions. That is just an argument for a lack of age-related generalization across studies, which virtually nullifies any theoretical basis this study might have ever had to consider its results meaningful in the first place. This is a common belief, but the evidence for it – popularly known as the Heckman Curve – is scant.[75] Regardless, if assessment of scaling is ever done, we can only hope that the analysis is well-powered, properly designed, and psychometrically sound.

Conclusion

Troller-Renfree et al. sought to assess how unconditional cash transfers affected brainwave activity in infants. The results they observed were null, but discussed as though they were positive, theory-confirming, and, most importantly, policy relevant. Their results were ambiguous, and, because their cited literature was so poorly powered and inconsistent, their theory formation should not be taken seriously.

The policy relevance of Troller-Renfree et al.'s null results was founded on the belief that they found effects; that those effects were beneficial; and that they are, or will be, large enough to generate positive behavioral and cognitive change in affected children. The evidence mustered in favor of the syllogism that would make these results meaningful was poor, contradicted by aspects of the study itself, and, at present, not worth serious consideration.

The contemporary obsession with the scientific justification of policy gives rise to politically motivated and shallow research, as well as the use of questionable research practices in the pursuit of a researcher's extra-scientific goals. University press departments and mainstream media will often credulously report such research, seemingly without any awareness of its weaknesses. The press departments won't raise a single criticism – they will write what they are told by a study's authors. The mainstream press tends to take the studies' authors too seriously when they have reached personally agreeable results, regardless of the quality of the work. Troller-Renfree et al. should serve as an example of low-quality review at every step, from peer review at PNAS to its reception by *The New York Times, Vox, Forbes*, and other outlets, and parts of the broader public.

It is difficult to say who is more at fault for popularizing what were essentially null results. Does the blame lie with the journalists, who had all the opportunity in the world to look at the research, read the citations, and realize it did not hold up? Or is it the fault of the authors, who pushed their paper through PNAS' easy track for submission and chose to promote an untenable interpretation of their findings to the wider world? It might be the fault of readers, who believed any of them. At every step, people acted irresponsibly. Scientists have a duty to correct misconceptions about their research, to prevent people from portraying it as more than it was, and to curb public excesses regarding the interpretation of their work.

The public arena is at least somewhat self-correcting, in that the findings in Troller-Renfree et al. were quickly critiqued by a handful of bloggers and scientists. The bigger problem, however, is that if it had not gotten a write up in the *New York Times* accompanied by a news alert, and if the study's topic and author's framing had not been so palpably absurd, the paper may have never been carefully checked and the results may have turned into conventional wisdom. Moreover, the handful of articles that criticized the study were likely seen by fewer people than the many articles that portrayed the study positively and uncritically. Bad science is the norm; correcting it is not.

Cases like this ought to make us reconsider the role social science plays in our public policy debates. Political bias and confirmation bias are heuristics that plague us all, and we have no reason to think that social scientists or policymakers are immune. Flashy results that support a popular policy are almost always untrustworthy, and large effects are usually exaggerated, p-hacked, or due to chance. The reality is that most social interventions and policies have a negligible impact when it comes to improving cognitive ability or behavior. Until researchers and the educated public come to grips with this fact, we should be skeptical of policymakers' ability to evaluate research objectively and social scientists' ability to reliably inform public policy with their work.

---

[1] R code for original analyses: https://rpubs.com/JLLJ/TTAN includes a number of power analyses and various tests related to this report; https://rpubs.com/JLLJ/MSRY includes methods to assess effect size overestimation and plots to aid in understanding the relationship between statistical power and overestimation; and https://rpubs.com/JLLJ/ODEN includes various adoption study-related plots.

[2] Elbel, Brian, Alyssa Moran, L Beth Dixon, Kamila Kiszko, Jonathan Cantor, Courtney Abrams, Tod Mijanovich. 2015. "Assessment of a Government-Subsidized Supermarket in a High-Need Area on Household Food Availability

and Children's Dietary Intakes." *Public Health Nutrition* 18(15): 2881-2890; Elbel, Brian, Tod Mijanovich, Kamila Kiszko, Courtney Abrams, Jonathan Cantor, L. Beth Dixon. 2017. "The Introduction of a Supermarket via Tax-Credits in a Low-Income Area: The Influence on Purchasing and Consumption." *American Journal of Public Health Promotion* 31(1): 59-66; Gittelsohn, Joel, Megan Rowan, Preety Gadhoke. 2012. "Interventions in Small Food Stores to Change the Food Environment, Improve Diet, and Reduce Risk of Chronic Disease." *Prev Chronic Dis 9.*

[3] The exception was from Shin, Ahyoung, Pamela J Surkan, Anastasia J Coutinho, Sonali R Suratkar, Rebecca K Campbell, Megan Rowan, Sangita Sharma, Lauren A Dennisuk, Micaela Karlsen, Anthony Gass and Joel Gittelsohn. 2015. "Impact of Baltimore Healthy Eating Zones: An Environmental Intervention to Improve Diet Among African American Youth." *Health Educ Behavior* 42(1): 97-105. They did not find effects on BMI in an experimental intervention. However, the authors claimed – improperly – to find an effect on obesity for the small and imbalanced overweight and obese female subgroup despite failing to test the significance of these claimed interaction effects. A result in their abstract was also portrayed as a BMI reduction resulting from their experiment, but that was a comparison between pre and post values for the intervention group, not between the treatment and control groups. Their intervention did not elicit any significant effects on BMI.

[4] Hurdle, Jon. 2010. "U.S. Launches Program to End 'Food Deserts.'" *Reuters*. Available at https://www.reuters.com/article/us-food-health-program/u-s-launches-program-to-end-food-deserts-idUSTRE61I5E820100219; "LYFTUP GROCERY ACCESS PROGRAM: Making Healthy Food Accessible." n.d. *Lyft*. Available at https://www.lyft.com/lyftup/grocery-access; "Healthy Communities." n.d. *Obama White House Archives*. Available at https://letsmove.obamawhitehouse.archives.gov/healthy-communities; Baily, Tres. 2014. "How We're Fighting Food Deserts." *Walmart*. Available at https://corporate.walmart.com/newsroom/health-wellness/20140519/how-were-fighting-food-deserts; "Search Results: 'Food Desert.'" n.d. *NIH Reporter*. Available at https://reporter.nih.gov/search/0OXiBizeRkOImyaZ-tzTlA/projects/charts.

[5] Muralidharan, Karthik and Abhijeet Singh. 2020. "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India." Working Paper. *National Bureau of Economic Research*;
West, Steven L. and Keri K. O'Neal. 2004. "Project D.A.R.E. Outcome Effectiveness Revisited." *American Journal of Public Health* 94(6): 1027-1029; Shackford, Stacey. 2010. "New Center, with $1 Million Grant, Aims to Make School Lunchrooms Smarter." *Cornell Chronicle*. Available at https://news.cornell.edu/stories/2010/10/cornell-gets-1-million-improve-school-nutrition. For a discussion of social psychology and behavioral science policies, see Singal, Jessie. 2021. "The Quick Fix: Why Fad Psychology Can't Cure Our Social Ills." Farrar, Straus and Giroux.

[6] DeParle, Jason. 2022. "Cash Aid to Poor Mothers Increases Brain Activity in Babies, Study Finds." *The New York Times.* Available at https://www.nytimes.com/2022/01/24/us/politics/child-tax-credit-brain-function.html.

[7] Troller-Renfree, Sonya V., Molly A. Costanzo, Greg J. Duncan, Katherine Magnuson, Lisa A. Gennetian, Hirokazu Yoshikawa, Sarah Halpern-Meekin, Nathan A. Fox and Kimberly G. Noble. 2022. "The Impact of a Poverty Reduction Intervention on Infant Brain Activity." *PNAS* 199(5).

[8] Matthews, Dylan. 2022. "Can Giving Parents Cash Help with Babies' Brain Development?" *Vox*. Available at https://www.vox.com/future-perfect/22893313/cash-babies-brain-development.

[9] Jirari, Tahra and Ed Prestera. 2022. "Cash Benefits to Low-Income Families May Aid Babies' Cognitive Development." *Niskanen Center*. Available at https://www.niskanencenter.org/cash-benefits-to-low-income-families-aids-babies-cognitive-development/; Columbia University. 2022. "Cash Support for Low-Income Families Impacts Infant Brain Activity." *Medical Xpress*. Available at https://medicalxpress.com/news/2022-01-cash-low-income-families-impacts-infant.html; Sullivan, Kaitlin. 2022. "Giving Low-Income Families Cash Can Help Babies' Brain Activity." *NBC News*. Available at https://www.nbcnews.com/health/health-news/poverty-hurts-early-brain-development-giving-families-cash-can-help-rcna13321; Smith, Zachary Snowden. 2022. "Giving Moms Money can Boost Babies' Brainwaves, Study Finds." *Forbes*. Available at https://www.forbes.com/sites/zacharysmith/2022/01/24/giving-moms-money-can-boost-babies-brain-activity-study-finds/?sh=2d2a456d20c7; for a full list of mentions and media appearances, see Troller-Renfree et al.'s press page for the study: https://www.babysfirstyears.com/press.

[10] "Press Release: Cash Support for Low-Income Families Impacts Infant Brain Activity." 2022. *Baby's First Years.* Available at https://www.babysfirstyears.com/press-release.

[11] Troller-Renfree et al. 2022. p. 5.

[12] Ritchie, Stuart [@StuartJRitchie]. 2022. *Great to see, part 2: [Tweet].* Twitter. https://twitter.com/StuartJRitchie/status/1486814686125375499; Ritchie, Stuart. 2022. "The Real Lesson of that Cash-for-Babies Study." *The Atlantic*. Available at https://www.theatlantic.com/science/archive/2022/02/cash-transfer-babies-study-neuroscience-hype/621488/.

[13] Gelman, Andrew. 2022. "I'm Skeptical of That Claim That "Cash Aid to Poor Mothers Increases Brain Activity in Babies." *Statistical Modeling, Causal Inference, and Social Science.* Available at https://statmodeling.stat.columbia.edu/2022/01/25/im-skeptical-of-that-claim-that-cash-aid-to-poor-mothers-increases-brain-activity-in-babies/.
[14] Alexander, Scott. 2022. "Against That Poverty and Infant EEGs Study." *Astral Codex Ten.* Available at https://astralcodexten.substack.com/p/against-that-poverty-and-infant-eegs?utm_source=url.
[15] Benasich, April A., Zhenkun Gou, Naseem Choudhury and Kenneth D. Harris. 2008. "Early Cognitive and Language Skills are Linked to Resting Frontal Gamma Power Across the First 3 Years." *Behavioural Brain Research* 195(2): 215-222; Gouam, Zhenkun, Naseem Choudhury and April A. Benasich. 2011. "Resting Frontal Gamma Power at 16, 24 and 36 Months Predicts Individual Differences in Language and Cognition at 4 and 5 Years." *Behavioural Brain Research* 220(2): 263-270; Brito, Natalie H., William P. Fifer, Michael M. Myers, Amy J. Elliott and Kimberly G. Noble. 2016. "Associations Among Family Socioeconomic Status, EEG Power at Birth, and Cognitive Skills During Infancy." *Developmental Cognitive Neuroscience* 19: 144-151*;* Maguire, Mandy J. and Julie M. Schneider. 2019. "Socioeconomic Status Related Differences in Resting State EEG Activity Correspond to Differences in Vocabulary and Working Memory in Grade School." *Brain and Cognition* 137; Williams, I. A., A. R. Tarullo, P. G. Grieve, A. Wilpers, E. F. Vignola, M. M. Myers and W. P. Fifer. 2012. "Fetal Cerebrovascular Resistance and Neonatal EEG Predict 18-Month Neurodevelopmental Outcome in Infants with Congenital Heart Disease." *Obstetrics & Gynaecology* 40(30): 304-309; Brito, Natalie H., Amy J. Elliott, Joseph R. Isler, Cynthia Rodriguez, Christa Friedrich, Lauren C. Shuffrey and William P. Fifer. 2019. "Neonatal EEG Linked to Individual Differences in Socioemotional Outcomes and Autism Risk in Toddlers." *Developmental Psychobiology* 61(8) 1110-1119.
[16] There are three exceptions. The first is their own study, the second is Marshall, Peter J., Bethany C. Reeb, Nathan A. Fox, Charles A. Nelson III and Charles H. Zeanah. 2008. "Effects of Early Intervention on EEG power and Coherence in Previously Institutionalized Children in Romania." *Development and Psychopathology,* and the third is Vanderwert, Ross E., Peter J. Marshall, Charles A. Nelson III, Charles H. Zeanah and Nathan A. Fox. 2010. "Timing of Intervention Affects Brain Electrical Activity in Children Exposed to Severe Psychosocial Neglect." *PLOS One.* However, the latter two studies used the same samples for the same sorts of analyses, and so it's plausible to think of them as a single exception. These were exceptions by nature of involving a randomized experiment.
[17] For a humorous illustration of the problem with multiple comparisons, see https://xkcd.com/882/.
[18] See Yarkoni, Tal. "Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009)." *Perspectives in Psychological Science* 4(3).
[19] Benasich et al. 2008; Harmony, Thalía, Erzsébet Marosi, Ana E. Díaz de León, Jacqueline Becker and Thalía Fernández. 1990. "Effect of Sex, Psychosocial Disadvantages and Biological Risk Factors on EEG Maturation." *Electroencephalography and Clinical Neurophysiology* 75(6): 482-91; Gou, Zhenkun, Naseem Choudhury and April A. Benasich. 2011. "Resting Frontal Gamma Power at 16, 24 and 36 Months Predicts Individual Differences in Language and Cognition at 4 and 5 Years." *Behavioural Brain Research* 220(2): 263-70; Williams, I. A., A. R. Tarullo, P. G. Grieve, A. Wilpers, E. F. Vignola, M. M. Myers and W. P. Fifer. 2012. "Fetal Cerebrovascular Resistance and Neonatal EEG Predict 18-Month Neurodevelopmental Outcome in Infants with Congenital Heart Disease." U*ltrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 40(3): 304-309; Brito et al. 2016; Brito et al. 2019; Cantiani, Chiara, Caterina Piazza, Giulia Mornati, Massimo Molteni and Valentina Riva. 2019. "Oscillatory Gamma Activity Mediates the Pathway from Socioeconomic Status to Language Acquisition in Infancy." *Infant Behavior and Development* 57; Troller-Renfree, Sonya V., Natalie H. Brito, Pooja M. Desai, Ana G. Leon-Santos, Cynthia A. Wiltshire, Summer N. Motton, Jerrold S. Meyer, Joseph Isler, William P. Fifer and Kimberly G. Noble. 2020. "Infants of Mothers with Higher Physiological Stress Show Alterations in Brain Function." *Developmental Science* 23(6); Maguire and Schneider 2019.
[20] Otero, Gloria. 1994. "EEG Spectral Analysis in Children with Sociocultural Handicaps." *The International Journal of Neuroscience* 79(3-4): 213-220; Otero, G. A., F. B. Pliego-Rivero, T. Fernández and J. Ricardo. 2003. "EEG Development in Children with Sociocultural Disadvantages: A Follow-up Study." *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology* 114(10): 1918-1925; McLaughlin, Katie A., Nathan A. Fox, Charles H. Zeanah, Margaret A. Sheridan, Peter Marshall and Charles A. Nelson. 2010. "Delayed Maturation in Brain Electrical Activity Partially Explains the Association Between Early Environmental Deprivation and Symptoms of Attention-Deficit/Hyperactivity Disorder." *Biological Psychiatry* 68(4): 329-336; Tomalski, Przemyslaw, Derek G. Moore, Helena Ribeiro, Emma L. Axelsson, Elizabeth Murphy, Annette

Karmiloff-Smith, Mark H. Johnson and Elena Kushnerenko. 2013. "Socioeconomic Status and Functional Brain Development - Associations in Early Infancy." *Developmental Science* 16(5): 676-687; Cantiani et al. 2019; Debnath, Ranjan, Alva Tang, Charles H. Zeanah, Charles A. Nelson and Nathan A. Fox. 2020. "The Long-Term Effects of Institutional Rearing, Foster Care Intervention and Disruptions in Care on Brain Electrical Activity in Adolescence." *Developmental Science* 23(1); Troller-Renfree et al. 2020; Brito, Natalie H., Sonya V. Troller-Renfree, Ana Leon-Santos, Joseph R. Isler, William P. Fifer and Kimberly G. Noble. 2020. "Associations Among the Home Language Environment and Neural Activity during Infancy." *Developmental Cognitive Neuroscience* 43; Jensen, Sarah K. G., Wanze Xie, Swapna Kumar, Rashidul Haque, William A. Petri and Charles A. Nelson. 2021. "Associations of Socioeconomic and Other Environmental Factors with Early Brain Development in Bangladeshi Infants and Children." *Developmental Cognitive Neuroscience* 50.

[21] A real-life and relevant example of effect size exaggeration comes from the literature on nerve conduction velocity (NCV). NCV was found to be positively related to IQ in samples that were generally small but ranged up to about 200 (Reed, Edward T., Philip A. Vernon and Andrew M. Johnson. 2004. "Confirmation of Correlation Between Brain Nerve Conduction Velocity and Intelligence Level in Normal Adults." *Intelligence* 32(6): 563-572.) The significant correlations in this literature were usually around 0.2-0.5 (see Vernon, Philip A. and Monica Mori. 1992. "Intelligence, Reaction Times, and Peripheral Nerve Conduction Velocity." *Intelligence* 16(3-4): 273-288; Budak, Faik, Tuncay Müge Filiz, Pinar Topsever and Üner Tan. 2009. "Correlations Between Nonverbal Intelligence and Nerve Conduction Velocities in Right-Handed Male and Female Subjects." *International Journal of Neuroscience* 115(5): 613-623; and Reed, T. Edward, Arthur R. Jensen. 1992. "Conduction Velocity in a Brain Nerve Pathway of Normal Adults Correlates with Intelligence Level." *Intelligence* 16(3-4): 259-272.) but often went up to between 0.6-0.7 (Tan, Üner. 1996. "Correlations Between Nonverbal Intelligence and Peripheral Nerve Conduction Velocity in Right-Handed Subjects: Sex-Related Differences." *International Journal of Psychophysiology* 22(1-2): 123-128.) But what happened when the association was examined in a sample of 4,462? It decreased to $r = 0.10$: less than half its significant values in the literature, with a sample size sufficient to detect a correlation of 0.06 with 99% power (Kirkegaard, Emil O. W., Michael A. Woodley and Helmuth S. Nyborg. 2017. "Nerve Conduction Velocity and Cognitive Ability: A Large Sample Study." *RPubs*. Available at https://rpubs.com/EmilOWK/NCV_IQ_VES.)

[22] They did not write that they assessed what effect imbalance had on their power, but they might have. Imbalanced sample sizes reduce power.

[23] Begus, Katarina and Elizabeth Bonawitz. 2020. "The Rhythm of Learning: Theta Oscillations as an Index of Active Learning in Infancy." *Developmental Cognitive Neuroscience* 45.

[24] Kovacs, Kristof and Andrew R. A. Conway. 2016. "Process Overlap Theory: A Unified Account of the General Factor of Intelligence." *Psychological Inquiry* 27(3).

[25] If you have three variables, X, Y, and Z, and you know that X and Y are positively correlated and Y and Z are positively correlated, but not whether X and Z are positively correlated, to know if X and Z are surely correlated *at some level above* $r = 0$, the correlations r_{xy} and r_{yz} can be squared and summed. If the resulting value exceeds 1, transitivity of associations is assured. This can appear to be contradicted in the real world if the correlations are estimated imprecisely.

[26] Posthuma, Daniëlle, Eco J. C. De Geus, Wim F. C. Baaré, Hilleke E. Hulshoff Pol, René S. Kahn and Dorret I. Boomsma. 2002. "The Association Between Brain Volume and Intelligence is Genetic in Origin." *Nature Neuroscience* 5: 83-84; See also Posthuma, Daniëlle, Wim F. C. Baaré, Hilleke E. Hulshoff Pol, René S. Kahn, Dorret I. Boomsma and Eco J. C. De Geus. 2012. "Genetic Correlations Between Brain Volumes and the WAIS-III Dimensions of Verbal Comprehension, Working Memory, Perceptual Organization, and Processing Speed." *Twin Research and Human Genetics* 6(2), Dreary, Ian J., Lars Penke and Wendy Johnson. 2010. "The Neuroscience of Human Intelligence Differences." *Nature Reviews Neuroscience* 11: 201-211, and Jansen, Philip R., Mats Nagel, Kyoko Watanabe, Yongbin Wei, Jeanne E. Savage, Christiaan A. de Leeuw, Martijn P. van den Heuvel, Sophie van der Sluis and Daniëlle Posthuma. 2020. "Genome-Wide Meta-Analysis of Brain Volume Identifies Genomic Loci and Genes Shared with Intelligence." *Nature Communications* 11. The environmental correlations generally run in the opposite direction, suggesting that environmental influences in aggregate act to reduce the relationships intelligence has with gray and white matter volumes.

[27] Plomin, Robert. 2014. "Genotype-Environment Correlation in the Era of DNA." *Behavior Genetics* 44: 629-638; Ericsson, Malin, Cecilia Lundholm, Stefan Fors, Anna K. Dahl Aslan, Catalina Zavala, Chandra A. Reynolds and Nancy L. Pedersen. 2017. "Childhood Social Class and Cognitive Aging in the Swedish Adoption/Twin Study of Aging." *PNAS*, 114(27): 7001-7006.

[28] Smit, D. J. A., D. Posthuma, D. I. Boomsma, E. J. C. De Geus. 2005. "Heritability of Background EEG Across the Power Spectrum." *Psychophysiology* 42(6): 691-697.
[29] Wax, Amy. 2017. "The Poverty of the Neuroscience of Poverty: Policy Payoff or False Promise?" *Faculty Scholarship at Penn Law*. Available at https://scholarship.law.upenn.edu/faculty_scholarship/1711/.
[30] Pierpont, Mary Ella, Martina Brueckner, Wendy K. Chung, Vidu Garg, Ronald V. Lacro, Amy L. McGuire, Seema Mital, James R. Priest, William T. Pu, Amy Roberts, Stephanie M. Ware, Bruce D. Gelb, Mark W. Russell and On behalf of the American Heart Association Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; and Council on Genomic and Precision Medicine. 2018. "Genetic Basis for Congenital Heart Disease: Revisited: A Scientific Statement from the American Heart Association." *Circulation* 138(21).
[31] Troller-Renfree et al. 2022. p. 7.
[32] Researchers preregister hypotheses prior to conducting statistical tests to improve the credibility of their results by emphasizing to others how strongly their hypotheses were tested and how their results were predicted rather than justified after the fact. This has also been described as a way to show how severe a test is (see Lakens, Daniel. 2019. "The Value of Preregistration for Psychological Science: A Conceptual Analysis." *JStage.* Available at https://www.jstage.jst.go.jp/article/sjpr/62/3/62_221/_article/-char/ja/).
[33] Link for the empty .zip file: https://ars.els-cdn.com/content/image/1-s2.0-S1878929320300281-mmc1.zip
[34] Matthews 2022.
[35] Yarkoni, Tal. "Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power—Commentary on Vul et al. (2009)." *Perspectives in Psychological Science.*
[36] Gelman, Andrew and Hal Stern. 2006. "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant." *The American Statistician* 60(4).
[37] See Nieuwenhuis, Sander, Birte U Forstmann and Eric-Jan Wagenmakers. 2011. "Erroneous Analyses of Interactions in Neuroscience: A Problem of Significance." *Nature Neuroscience* 14: 1105-1107.
[38] Gelman, Andrew. 2018. "You Need 16 Times the Sample Size to Estimate an Interaction Than to Estimate a Main Effect." *Statistical Modeling, Causal Inference, and Social Science*. Available at https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/.
[39] Miller, William R. and Martha Sanchez-Craig. 1996. "How to Have a High Success Rate in Treatment: Advice for Evaluators of Alcoholism Programs." *Addiction* 91(6): 779-785.
[40] Simmons, Joseph P., Leif D. Nelson and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11).
[41] Wicherts, Jelte M., Coosje L. S. Veldkamp, Hilde E. M. Augusteijn, Marjan Bakker, Robbie C. M. van Aert and Marcel A. L. M. van Assen. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking." *Frontiers in Psychology* 25.
[42] "ASQ-3." n.d. *ASQ: Ages and Stages Questionnaire.* Available at https://agesandstages.com/products-pricing/asq3/.
[43] The authors were clearly aware of this method based on their preregistration where they said, "all measures will be examined for psychometric equivalence across race/ethnicity and whether Spanish or English is a primary language spoken at home," but they did not apply them to treatment effects.
[44] Protzko, John, Jan te Nijenhuis, Khaled Ziada, Hanaa Abdelazim, Mohamed Metwaly and Salaheldin Bakhiet. 2021. "What to do Without a Control Group: You Have to go Latent, but not all Latents are Equal." *PsyArXiv.* Available at https://psyarxiv.com/vymp3/.
[45] Matthews 2022.
[46] Protzko, John. 2015. "The Environment in Raising Early Intelligence: A Meta-Analysis of the Fadeout Effect." *Intelligence* 53: 202–210.
[47] Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman and David S. Yeager. 2020. "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions." *Psychological Science in the Public Interest* 21(2).
[48] Though it has also been observed for many phenotypes and its name was given based on it first being observed for height and weight (see Wilson, Ronald S. 1976. "Concordance in Physical Growth for Monozygotic and Dizygotic Twins." *Anatomy of Human Biology* 3(1)).
[49] Bouchard, Thomas J. Jr. 2013. "The Wilson Effect: The Increase in Heritability of IQ With Age." *Twin Research and Human Genetics* 16(5).

[50] DeParle 2022.
[51] During the writing process for this manuscript, the .pdf cited here was deleted, but I have archived it: https://web.archive.org/web/20210309033815/https://www.pnas.org/sites/default/files/advanced-pages/reviewprocess.pdf.
[52] For a smattering of comments on PNAS' Contributions, see Rand, David G. and Thomas Pfeiffer. 2009. "Systematic Differences in Impact across Publication Tracks at PNAS." *PLOS One*; Davis, Philip M. 2016. "Comparing the Citation Performance of PNAS Papers by Submission Track." *bioRxiv*; Aldhous, Peter. 2014. "Scientific Publishing: The Inside Track." *Nature* 510: 330-332; "OPENING UP PEER REVIEW: THE PECULIAR CASE OF PNAS CONTRIBUTED PAPERS." 2016. *Rapha-Z-Lab.* Available at https://raphazlab.wordpress.com/2016/01/19/opening-up-peer-review-the-peculiar-case-of-pnas-contributed-papers/; Lowe, Derek. 2008. "PNAS: Read It, or Not?" *Science*. Available at https://www.science.org/content/blog-post/pnas-read-not.
[53] These articles were Noble, Kimberly G. and Martha J. Farah. 2013. "Neurocognitive Consequences of Socioeconomic Disparities: The Intersection of Cognitive Neuroscience and Public Health." *Developmental Science* 16(5): 639-640; Noble, Kimberly G., Bruce D. McCandliss and Martha J. Farah. 2007. "Socioeconomic Gradients Predict Individual Differences in Neurocognitive Abilities." *Developmental Science* 10(4): 464-480; Romer, Daniel and Elaine F. Walker. 2007. "Adolescent Psychopathology and the Developing Brain: Integrating Brain and Prevention Science." *Oxford Scholarship Online;* Noble, Kimberly G., Martha J. Farah and Bruce D. McCandliss. 2006. "Socioeconomic Background Modulates Cognition–Achievement Relationships in Reading Author Links Open Overlay Panel." *Cognitive Development* 21(3): 349-368; Noble, Kimberly G., Michael E. Wolmetz, Lisa G. Ochs, Martha J. Farah and Bruce D. McCandliss. 2006. "Brain–Behavior Relationships in Reading Acquisition are Modulated by Socioeconomic Factors." *Developmental Science* 9(6): 642-654; Noble, Kimberly G., M. Frank Norman and Martha J. Farah. 2004. "Neurocognitive Correlates of Socioeconomic Status in Kindergarten Children." *Developmental Science* 8(1): 74-87.
[54] Dr. Luby may also have been an unethical reviewer choice because of potential prior ethical abuses on her part. See here: https://www.cbsnews.com/news/updated-doc-who-urged-antipsychotics-for-3-year-olds-funded-by-j038j-az-and-shire/.
[55] Sylvester, Chad M., Deanna M. Barch, Michael P. Harms, Joan L. Luby, Nathan A. Fox and Daniel S. Pine. 2015. "Early Childhood Behavioral Inhibition Predicts Cortical Thickness in Adulthood." *J Am Acad Child Adolesc Psychiatry* 55(2).
[56] Kraft, Matthew A. 2019. "Interpreting Effect Sizes of Education Interventions." *Annenberg Institute at Brown University.* Available at https://scholar.harvard.edu/files/mkraft/files/kraft_2019_effect_sizes.pdf.
[57] "Head Start Impact Study: Final Report, Executive Summary." 2010. *Office of Planning, Research & Evaluation.* Available at https://www.acf.hhs.gov/opre/report/head-start-impact-study-final-report-executive-summary.
[58] Lortie-Forgues, Hugues and Matthew Inglis. 2019. "Rigorous Large-Scale Educational RCTs Are Often Uninformative: Should We Be Concerned?" *Educational Researcher* 48(3).
[59] Publication bias is the phenomenon whereby significant or otherwise favorable results are more likely to be published. It is often evidenced, for example, by correlating the standard errors of studies with their point estimates or by assessing the symmetry of point estimates around the meta-analytic mean. The first is informative because at some level, smaller studies must yield larger effects, and the pattern also suggests searching for significance because researchers want significant results, but their sample sizes are constrained. This latter suggestion brings us to the second method of checking for asymmetry of points around the meta-analytic mean. Asymmetry means that there are omitted studies with estimates on one side of the meta-analytic mean but can also indicate that larger studies yield smaller effects. This pattern could be argued to emerge because of, for example, smaller studies in a meta-analysis of experiments being more intensive because they can dedicate their limited resources to a smaller group, whereas larger studies must spread out their resources over more people, diluting possible effects. While tempting, the fact that this pattern frequently emerges in studies of experimental and nonexperimental (e.g., correlational) research suggests it is *not* due to differences in the intensiveness of programs. Suggestions to that effect as explanations in particular cases need to be investigated and evidenced rigorously, Moreover, it is often larger programs that are more intensive rather than the reverse. Because of the ubiquity of patterns indicative of publication bias regardless of the experimental vs. nonexperimental nature of studies, many reasonably take it for granted that it is evidence of bias. As a final note, publication bias can become prevalent for anodyne reasons. For example, if an early study in a given literature erroneously reports a large effect, subsequent studies may seek to confirm its results and may perform power analyses with its effect size in mind, leading to underestimated requisite sample sizes and, due to the search for significance, the need to themselves produce overestimates. If replication

attempts are conducted by different teams whose power analyses are conducted similarly – which is not unlikely – then those teams who achieve extreme results may be more likely to publish than teams whose results are not as large and significant. Importantly, in the social scientific world, where uncertainty about point estimates is the norm, it is not appropriate to forego these power analyses or to use exceptionally high effect sizes in them because one claims to know about how large an effect should be prior to the publication of analyses that are well-powered to detect small effects, barring some other exceptional reasoning. This example of persistently inflated effects resulting from a need for significance and trust in erroneous but early results is comparable to Feynman's famous cargo cult example involving Millikan's oil-drop experiment.

[60] Duncan, Greg J. and Katherine Magnuson. 2013. "Investing in Preschool Programs." *Journal of Economic Perspectives* 27(2): 109-132.

[61] Whitehurst, Grover J. "Russ." 2018. "Does State Pre-K Improve Children's Achievement?" *Brookings Institution.* Available at https://www.brookings.edu/research/does-state-pre-k-improve-childrens-achievement/.

[62] Arriagada, Ana-Marie, Jonathan Perry, Laura Rawlings, Julieta Trias and Melissa Zumaeta. 2018. "Promoting Early Child Development through Combining Cash Transfers and Parenting Programs." *World Bank Group.* Available at https://documents1.worldbank.org/curated/en/827231544474543725/pdf/WPS8670.pdf.

[63] Dulal, Sophiya, Frédérique Liégeois, David Osrin, Adam Kuczynski, Dharma S. Manandhar, Bhim P. Shrestha, Aman Sen, Naomi Saville, Delan Devakumar and Audrey Prost. 2018. "Does Antenatal Micronutrient Supplementation Improve Children's Cognitive Function? Evidence from the Follow-Up of a Double-Blind Randomised Controlled Trial in Nepal." *BMJ Global Health* 3.

[64] Behrens, Annika, Elmar Graessel, Anna Pendergrass and Carolin Donath. 2020. "Vitamin B—Can it Prevent Cognitive Decline? A Systematic Review and Meta-Analysis." *Systematic Reviews* 9.

[65] Ali, Hasmot, Jena Hamadani, Sucheta Mehra, Fahmida Tofail, Md Imrul Hasan, Saijuddin Shaikh, Abu Ahmed Shamim, Lee S-F Wu, Keith P. West, Jr. and Parul Christian. 2017. "Effect of Maternal Antenatal and Newborn Supplementation with Vitamin A on Cognitive Development of School-Aged Children in Rural Bangladesh: A Follow-Up of a Placebo-Controlled, Randomized Trial." *The American Journal of Clinical Nutrition* 106(1): 77-87.

[66] Welch, Vivian A., Elizabeth Ghogomu, Alomgir Hossain, Alison Riddle, Michelle Gaffey, Paul Arora, Omar Dewidar, Rehana Salam, Simon Cousens, Robert Black, T. Déirdre Hollingsworth, Sue Horton, Peter Tugwell, Donald Bundy, Mary Christine Castro, Alison Elliott, Henrik Friis, Huong T. Le, Chengfang Liu, Emily K. Rousham, Fabian Rohner, Charles King, Erliyani Sartono, Taniawati Supali, Peter Steinmann, Emily Webb, Franck Wieringa, Pattanee Winnichagoon, Maria Yazdanbakhsh, Zulfiqar A. Bhutta and George Wells. 2019. "Mass Deworming for Improving Health and Cognition of Children in Endemic Helminth Areas: A Systematic Review and Individual Participant Data Network Meta-Analysis." *Systematic Review* 15(4).

[67] Taylor-Robinson, David C., Nicola Maayan, Karla Soares-Weiser, Sarah Donegan and Paul Garner. 2015. "Deworming Drugs for Soil-Transmitted Intestinal Worms in Children: Effects on Nutritional Indicators, Haemoglobin, and School Performance." *Cochrane Database of Systematic Reviews* 7; Welch, Vivian A., Elizabeth Ghogomu, Alomgir Hossain, Shally Awasthi, Zulfi Bhutta, Chisa Cumberbatch, Robert Fletcher, Jessie McGowan, Shari Krishnaratne, Elizabeth Kristjansson, Salim Sohani, Shalini Suresh, Peter Tugwell, Howard White and George Wells. 2016. "Deworming and Adjuvant Interventions for Improving the Developmental Health and Well-Being of Children in Low- And Middle-Income Countries: A Systematic Review and Network Meta-Analysis." *Campbell Systematic Reviews* 7.

[68] For more information on the Dual N-Back, see this FAQ: https://www.gwern.net/DNB-FAQ.

[69] "Dual N-Back Meta-Analysis." 2018. *Gwern.net*. Available at https://www.gwern.net/DNB-meta-analysis.

[70] Sala, Giovanni, Deniz N. Aksayli, Semir K. Tatlidil, Tomoko Tatsumi, Yasuyuki Gondo and Fernand Gobet. 2019. "Near and Far Transfer in Cognitive Training: A Second-Order Meta-Analysis." *Collabra: Psychology* 5(1): 18.

[71] Ludwig, Jens, Greg J. Duncan, Lisa A. Gennetian, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling and Lisa Sanbonmatsu. 2013. "Long-Term Neighborhood Effects on Low-Income Families: Evidence from Moving to Opportunity." *American Economic Review,* 103(3): 226-231.

[72] The modesty of adoption effects, which typically entail massive improvements in family socioeconomic status in all dimensions, is part of why heavily adjusted effects like those presented in Duncan, G. J., Morris, P. A. and Rodrigues, C. 2011. "Does Money Really Matter? Estimating Impacts of Family Income on Young Children's Achievement with Data from Random-Assignment Experiments." *Developmental Psychology* 47(5): 1263–1279 for much smaller improvements are implausible. The same is true for Morris, P., Duncan, G. J. and Clark-Kauffman, E. 2005. "Child Well-Being in an Era of Welfare Reform: The Sensitivity of Transitions in Development to Policy Change." *Developmental Psychology* 41(6): 919–932.

[73] A prime example of one of the largest adoption studies yielding modest effects is Kendler, Kenneth S., Eric Turkheimer, Henrik Ohlsson, Jan Sundquist and Kristina Sundquist. 2015. "Family Environment and the Malleability of Cognitive Ability: A Swedish National Home-Reared and Adopted-Away Cosibling Control Study." *PNAS*, 112(15): 4612-4617; See also Ericsson et al., 2017.
[74] Odenstad, A., A. Hjern, F. Lindblad, F. Rasmussen, B. Vinnerljung and M. Dalen. 2008. "Does Age at Adoption and Geographic Origin Matter? A National Cohort Study of Cognitive Test Performance in Adult Inter-Country Adoptees." *Psychological Medicine* 38(12).
[75] Gelman, Andrew. 2020. "Heckman Curve Update Update." Statistical Modeling, Causal Inference, and Social Science. Available at https://statmodeling.stat.columbia.edu/2020/08/12/heckman-curve-update-update/.