Free Associations: Psychoanalysis and Culture, Media, Groups, Politics
Number 89, Fall 2023 ISSN: 2047-0622
URL: http://www.freeassociations.org.uk/

The 1990 "Minnesota Study of Twins Reared-Apart" IQ Study: Ripe for Retraction?

Jay Joseph

Reared-apart (separated) identical twins have great appeal. Though exceedingly rare, many people see them as providing the ultimate Rosetta Stone-type method of teasing apart the potential roles of nature and nurture as causes of human behavioral variation. They share a genetic identity while—in theory—growing up and living under entirely different environmental conditions. The two main ways they have come to our attention have been anecdotal stories of individual twin pairs said to share "spooky" and "eerie" behavioral similarities, and a small handful of scientific studies based on a sample of twins. Here, I focus on the latter.

The "Minnesota Study of Twins Reared Apart" (MISTRA) is one of the most famous and widely cited studies in the behavioral sciences. The study began in 1979 and ended in 2000. Many academic publications based on the MISTRA data have appeared since 2000. The MISTRA IQ study was key, due to cognitive ability's long-standing central role in the "nature-nurture debate" and the fields of psychology and behavioral genetics. The study was published in a 1990 edition of *Science*, one of the world's top scientific journals. Although the researchers assessed other psychological characteristics in that article, I refer to it here as the "IQ study" because this was the main MISTRA IQ publication.

The 1990 MISTRA sample consisted of 56 reared-apart MZ twin pairs (monozygotic, identical; 100% genetic similarity) and 30 reared-apart DZ twin pairs (dizygotic, fraternal; average 50% genetic similarity). The three MISTRA cognitive ability (IQ) measures were the "Wechsler Adult Intelligence Scale" (WAIS), the "Raven's Progressive Matrices/Mill-Hill Vocabulary Scale composite," and the "First Principal Component of Special Mental Abilities" (FPC). Reared-apart MZ twins are known as "MZA" pairs; reared-apart DZ twins are known as "DZA" pairs.

Study initiator psychologist Thomas J. Bouchard, Jr., along with David Lykken, Matthew McGue, Nancy Segal, and Auke Tellegen concluded in the 1990 *Science* article that their twin study results showed that "IQ is strongly affected by genetic factors." They estimated the heritability of IQ at "about 70%." (The use of heritability estimates in behavioral research, as well as IQ testing itself, have been disputed for decades.) MISTRA IQ heritability estimates have been cited ever since in psychology textbooks, review publications, media reports, in the social media, and in books such as the controversial 1994 *The Bell Curve*. The *Science* article has been cited over 2,300 times since 1990 (about 70 citations per year).

My 2022 Analysis and Bouchard's 2023 Response

I have been writing about the MISTRA and other "twins reared apart" (TRA) studies for many years. In 2022, I published a newly formulated critical analysis in the journal *Human Development* (pdf available online, Volume 66, Issue 1). I concluded that contrary to most of what has been written about it, the "MISTRA IQ study failed to discover evidence that genetic factors influence IQ scores and cognitive ability across the studied population," and that the study was a replication-crisis exemplar of flawed science. I review and elaborate on these points below. The *replication crisis* has called into question the integrity of research in the behavioral sciences and science as a whole, and has been described in psychology as follows:

> "The replication crisis in psychology refers to concerns about the credibility of findings in psychological science. The term, which originated in the early 2010s, denotes that findings in behavioral science often cannot be replicated: Researchers do not obtain results comparable to the original, peer-reviewed study when repeating that study using similar procedures. For this reason, many scientists question the accuracy of published findings and now call for increased scrutiny of research practices in psychology."

In 2023, the now-retired Bouchard published a response to my article in the journal *Twin Research and Human Genetics* (published online on 6/5/2023), where he claimed to have refuted my central arguments. In the present review, I respond directly to Bouchard's 2023 article and examine the question of whether the 1990 MISTRA IQ study should be added to the growing list of retracted scientific research publications. I limit my response to the most important areas of contention.

Although Bouchard concluded in 2023 that the MISTRA findings are valid in part because they should be evaluated in the context of other studies (more on this point later), he did not cite any discoveries of genes for IQ or cognitive ability at the molecular genetic level. Bouchard had recognized in 2014 that the results of gene searches beginning in the early 1990s "have been dismal in comparison with expectation," and his failure to claim gene associations or discoveries in 2023 suggests that he continues to hold this view.

The key points I raised in my 2022 article were as follows:

1) "Twins reared apart" (TRA) studies contain numerous non-genetic similarity-producing biases that critics have documented for over 50 years, which call into question claims that environmental confounds are minor or absent in TRA studies.
2) The MISTRA design reproduced most of these biases, including that its volunteer twin participants were not separated at birth and randomly assigned to available adoptive homes, and that most pairs grew up knowing they had a twin sibling and having had contact with each other.
3) Most pairs found in TRA studies, including the MISTRA pairs, were only *partially* reared apart.

4) DZAs (DZ or fraternal twins reared apart) were the official MISTRA control group. The MISTRA was the first TRA study to recruit DZAs as controls.
5) The MISTRA researchers failed to publish the DZA control group IQ correlations in their 1990 *Science* article, and the full-sample DZA IQ correlations remain unpublished to this day.
6) After Bouchard and colleagues omitted and bypassed their 1990 DZA control group correlations, most likely to avoid a 0% heritability finding, they used their MZA (MZ or identical twins reared apart) IQ correlations alone to estimate heritability, based on their assumption that the MZA correlation "directly estimates heritability."
7) The assumption that the MZA IQ correlation directly estimates heritability is false. Even perfectly separated MZ twin pairs share many similarity-producing environmental and "cohort" influences in common. The researchers, on the other hand, assumed that similarity-producing environmental influences do not exist. They maintained this assumption by deciding to count most environmental influences as genetic influences.
8) Near-full-sample DZA IQ correlations published much later, in 2007 and 2012, show that the MISTRA MZA group and DZA group IQ correlations do not differ at a statistically significant level. This finding, by itself, leads to a conclusion that the IQ study found no evidence that hereditary factors influence IQ scores (0% IQ heritability).
9) When TRA study researchers fail to make their unpublished raw data and twins' life histories available for inspection and analysis by qualified independent reviewers, we must evaluate their findings with extreme caution—or even reject them entirely— because these findings are based on virtually irreproducible sample populations.
10) The MISTRA should be evaluated in the context of science's replication crisis, and the researchers used at least two *p-hacking* methods to arrive at a finding of above-zero IQ heritability. Strong genetic confirmation biases influenced their interpretations of the data.

A Closer Look at Bouchard's 2023 Response

Design of the Study. In my article, I quoted psychologist Raymond Fancher's description of a "definitive" or "ideal" TRA study, which would be based on randomly assigned and completely separated twin pairs representative of the total population of reared-apart twins. I concluded, "the MISTRA did not come close to meeting this standard." Bouchard responded by explaining the difference between a true or "planned experiment" and a "natural experiment" such as the MISTRA:

> "We did not conduct a true experiment. We gathered a sample of convenience and justified our conclusions on the grounds that the sample was reasonable for our purposes. The same is true for all 'natural experiments.'"

I'm not opposed to natural experiments. However, I pointed out that because researchers conducting them are unable to design, control, observe, or manipulate environmental conditions and other variables, they must make *assumptions* about these conditions and variables, and these assumptions must be valid. Critics argue that many TRA study assumptions are not valid, and I quoted McGue and Bouchard's own 1989 recognition that "several" MISTRA model-fitting assumptions "are likely not to hold for cognitive abilities." Bouchard did not mention or correct this little-known 1989 statement.

Bouchard mistakenly said that Fancher described a "perfect" study, which set up his later comment that "no study is perfect, including MISTRA." Everyone understands that studies of human behavior are not perfect. Still, I quoted Fancher to show that the MISTRA and other TRA studies did not meet minimum design requirements.

How "Separated" Were the MISTRA MZA Pairs? Five IQ studies based on supposedly "reared-apart" MZ twins (MZAs) have been published. British psychologist Cyril Burt's IQ TRA study was discredited in the 1970s on suspicions of fraud, now established "beyond a reasonable doubt," and Burt's publications are no longer part of the TRA IQ literature. The authors of the first three studies, Horatio Newman and colleagues in 1937, James Shields in 1962, and Niels Juel-Nielsen in 1965/1980 provided detailed case history information for most of the pairs they studied. The MISTRA and the authors of the final study, the Swedish Adoption/Twin Study on Aging (SATSA), produced only a few case histories. These case histories consisted mainly of cherry-picked MISTRA pairs, such as the "Jim Twins," that have been reported repeatedly in the media and MISTRA publications for decades. (For more on the genetically misleading aspects of these stories, see my review of the movie *Three Identical Strangers*.)

I described TRA studies and their problems in detail in my 2015 book *The Trouble with Twin Studies*. In my 2022 article, I showed that based on my analysis of the Newman, Shields, and Juel-Nielsen case histories:

- In 25/75 (33%) of the pairs, twins were separated at 12 months of age or later.
- In 56/75 (75%) of the pairs, twins had contact with each other while growing up.
- In 42/75 (56%) of the pairs, one or both twins were placed with a family member.
- In 17/75 (23%) of the pairs, twins lived together for at least 12 months after separation, or grew up next door to each other.

Bouchard didn't dispute these calculations or challenge my conclusion that "by default and until proven otherwise we must assume that the MISTRA MZA pairs were no more 'reared apart' than were…the partially reared-apart MZA samples found in the original three TRA studies." Like the earlier investigations, his was a study of twins only *partially* reared apart.

Failed Attempt to Invalidate Earlier Critics of TRA Research. As he has done since the 1980s, Bouchard spent about a quarter of his 2023 article attempting to discredit critics of the earlier studies, such as psychologist Leon Kamin in his pioneering 1974 book *The Science and Politics of I.Q.*, sociologist Howard Taylor's 1980 *The IQ Game*, and psychologist Susan Farber's 1981 *Identical Twins Reared Apart: A Reanalysis*. These books appeared in the aftermath of psychologist Arthur Jensen's highly publicized promotion of TRA research in 1969 and 1970 in support of his claims in favor of high within-group IQ heritability, and at least partial genetic causes of mean IQ differences between ethnic groups. Jensen relied heavily on the subsequently discredited Burt data (discussed in my 2018 tribute to Kamin).

In the 1980s and his 2023 article, Bouchard called the works of Kamin, Taylor, and Farber "pseudoanalyses" due to these authors' use of subgroup analyses in support of the importance of environmental influences on MZA behavioral and IQ resemblance. As one example of a subgroup analysis, Taylor showed that MZA pairs in the first three investigations who had been reunited before being studied correlated significantly higher for IQ than MZA pairs who had not been reunited. Bouchard claimed that he refuted these subgroup analyses long ago. Whether he succeeded in doing so is a topic for another article.

Although Bouchard's "walking in the garden of forking paths" data analysis description might have some relevance to the subgroup analyses he criticized, it has no relevance to the arguments I presented in my article. Like Bouchard, I am not a fan of using post-publication subgroup analyses to argue in favor of TRA study environmental biases, and in my 2022 article I did not perform or support such analyses. In attempting to knock over the subgroup analysis straw man he created, Bouchard tried to link me to a practice I did not engage in, depend on, or endorse.

Bouchard referred to Kamin, Taylor, and Farber as "discredited sources," and he attempted to paint me with the same brush by alleging that I "depended on" and "repeatedly cited" their "discredited" arguments. These were top-notch analysts whose work remains of great importance. They are not discredited sources just because Bouchard says they are, or for any other reason. He scolded journalist John Horgan for writing 30 years ago, "Kamin has shown that identical twins supposedly raised apart are often raised by members of their families or by unrelated families in the same neighborhood; some twins had extensive contact with each other while growing up." Was Bouchard implying that Kamin's descriptions as cited by Horgan were untrue, or that these factors do not bias TRA study results? Kamin's descriptions were accurate, as anyone who has read the first three studies' case descriptions can confirm.

My article briefly summarized the main TRA study problems and biases described by Kamin, Taylor, Farber, and others, most of which also apply to the MISTRA studies. These problems and biases related to the lack of separation described above, how twins were recruited to the studies, how they were studied, the environmental similarities they

experienced, and how they were evaluated and reported. Subgroup analyses aside, all these points remain in full force.

Suppression of the MISTRA DZA Control Group Data. According to Bouchard, "Joseph implies that we concealed data gathered from the dizygotic twins reared apart (DZAs)." I not only implied it, I said so directly. I continue to do so here.

The MISTRA was the first TRA study to recruit DZA pairs as controls. In my article, I argued that Bouchard and colleagues suppressed (omitted and bypassed) their DZA control group data to arrive at desired conclusions. This allegation is serious, and I do not make such allegations lightly. Yet in his 2023 article Bouchard didn't dispute my contentions (1) that DZAs were the official MISTRA control group, (2) that in the 1990 *Science* article he and his colleagues omitted their DZA correlations and sidestepped the MISTRA "model-fitting" procedure, (3) that the *Science* article reported no control group DZA results of any kind, and (4) that the MISTRA full-sample DZA IQ correlations were never published.

My article's Table 1 showed the reported 1990 *Science* correlations and noted the *non-reported* DZA control group correlations. The three MZA IQ samples ranged from 42-48 pairs.

Table 1. Reported and nonreported correlations in the 1990 MISTRA *Science* Article

Measure	MZA	MZT	DZA control group
WAIS full-scale IQ	*0.69*	*0.88*	*not reported*
Raven, Mill-Hill composite IQ	*0.78*	*0.76*	*not reported*
First principal component IQ	*0.78*	*"not available"*	*not reported*
Hawaii Special Mental Abilities	0.45	"not available"	not reported
Comprehensive Special Mental Abilities	0.48	"not available"	not reported
Multidimensional Personality Questionnaire	0.50	0.49	not reported
California Psychological Inventory (personality)	0.48	0.49	not reported
Strong Campbell Interest Inventory	0.39	0.48	not reported
Religiosity Scales	0.49	0.51	not reported
MPQ Traditionalism Scale	0.53	0.50	not reported

IQ measures are italicized. Intraclass correlations from Bouchard et al. (1990a, p. 226), Table 4. DZA, DZ twins reared apart; MZA, MZ twins reared apart; MZT, MZ twins reared together; WAIS, Wechsler Adult Intelligence Scale; MPQ, Multidimensional Personality Questionnaire. The 1990 MISTRA *Science* sample consisted of 56 MZA and 30 DZA pairs (p. 223). The MISTRA-reported MZA IQ correlations were based on fewer pairs, ranging from 42 to 48 pairs. "Not available" status was reported by the researchers. No DZA correlations or results of any kind were reported in the 1990 MISTRA *Science* article.

As in 1990, Bouchard wrote in 2023 that the 30 DZA pairs in the 1990 IQ study "were not included because the sample was small." He made this decision arbitrarily, presumably after reviewing the data. My article cited examples of *non-IQ* MISTRA studies appearing around the same time, where Bouchard published full-sample DZA correlations based on similar or even smaller DZA sample sizes (example here). I also showed that in her 2012 book *Born Together—Reared Apart: The Landmark Minnesota Twin Study*, MISTRA researcher and 1990 *Science* article co-author Nancy Segal revealed

that the researchers submitted an early-1980s paper to *Science* with IQ results based on only *twelve* DZA pairs. Bouchard did not dispute or comment upon any of the above examples.

I then provided Table 2, which included *near*-full-sample MISTRA DZA correlations published in 2007 and 2012 showing that the MISTRA IQ *r*MZA (the MZA group IQ correlation) and IQ *r*DZA (the DZA group IQ correlation) did not differ at a statistically significant level. If *r*MZA and *r*DZA do not differ significantly, we can safely conclude that non-genetic factors alone raised both IQ correlations above zero (more on this point below). As seen in Table 2, MZAs' greater genetic resemblance (100%) did not lead to their greater IQ behavioral resemblance versus DZAs (50% average genetic similarity; links to the Table 2 Johnson et al. 2007 and Segal 2012 data sources here and here).

Table 2. Near-full-sample MISTRA IQ correlations at the study's end: MZA versus DZA twin pairs

	MZA pairs (experimental group)	DZA pairs (control group)	Probability value
Wechsler (WAIS) IQ correlations	0.62 (74 pairs)	0.50 (52 pairs)	$p = 0.17$ Not statistically significant at the 0.05 level
Raven's Progressive Matrices IQ correlations	0.55 (74 pairs)	0.42 (52 pairs)	$p = 0.18$ Not statistically significant at the 0.05 level

Intraclass correlations. One-tailed probability. The final 2000 MISTRA full sample consisted of 81 MZA and 56 DZA pairs. Based on calculations made at the VassarStats website (http://vassarstats.net/rdiff.html). DZA, DZ twins reared apart; MZA, MZ twins reared apart; *p*, one-tailed probability; WAIS, Wechsler Adult Intelligence Scale. Sources: Wechsler (WAIS) correlations from Segal (2012, p. 286), based on the number of pairs reported on p. 284; Raven correlations from Johnson et al. (2007, p. 552), based on the number of pairs reported on p. 545. The DZA sample contained 18 opposite-sex pairs (Segal, 2012, p. 42). The DZA group correlation for the MISTRA "First Principal Component of Special Mental Abilities" measure, which was the third of three MISTRA IQ measures, has never been published.

Failure at the "Important First Step." In *Born Together—Reared Apart*, Segal emphasized that the MZA-DZA comparison is "an important first step" in determining "whether or not" genes influence IQ and other behavioral characteristics. In my article, I quoted Segal:

> "The simple comparison of the MZ (or MZA) and DZ (or DZA) intraclass correlations *is an important first step* in behavioral-genetic analysis because this demonstrates *whether or not* there is genetic influence on the trait" (emphasis added).

And elsewhere in the book, Segal wrote,

> "Genetic effects are shown *if* the correlation for MZ or MZA twins exceeds the correlation for DZ or DZA twins" (emphasis added).

The Swedish (SATSA) researchers agreed with Segal. "When MZ correlations are not greater than DZ correlations," they wrote in 1992, "twin similarity may reflect correlated environments rather than genetic similarity." Segal described a TRA study process in which intraclass "correlations are calculated separately for MZA and DZA twin pairs and compared." Yet in the MISTRA IQ study, MZA and DZA twin correlations *were not*

compared. I showed in 2022 (Table 2) that the MISTRA IQ study *failed* at its "important first step," and for this reason alone the study found no evidence that genes influence IQ (0% heritability).

Ironically, Bouchard criticized me for supposedly "walking in the garden of forking paths." Did he notice my 2022 Figure 1, entitled "The MISTRA: Two Paths to Genetic Findings"? There, I suggested that he and his colleagues arrived at their own 1990 "forking path" study decision point. I diagrammed and described how after discovering that the planned MISTRA IQ heritability path was blocked by the statistically non-significant "important first step" MZA-DZA comparison, they decided to use an IQ heritability path based only on the MZA results.

"Small Sample" and "Space Limitations." In addition to the supposedly small size of the DZA group, Bouchard said in the 1990 article that he could not publish his DZA correlations due to "space limitations." He wrote in 2023 that the purpose of the 1990 IQ study "was to report a constructive replication of previous studies of MZA twins in the brief format provided by *Science* and explain the methodology underlying the study of MZA twins." The 1990 article ran over 6,000 words (six pages) and therefore was not "brief," and the term "constructive replication" did not appear in it.

Bouchard didn't mention or dispute the numbers I presented in Table 2 or provide an acceptable explanation for why his control group DZA correlations did not appear in the 1990 *Science* article. Nor did he comment on my quoting Segal's description of the MISTRA "important first step" MZA-DZA comparison, or say that Segal was wrong.

Replicating or Overturning Previous TRA Studies? For Bouchard, the MISTRA IQ findings "replicated" those of the earlier three TRA studies. *In fact, they overturned them*. As I and others have shown, most MZA pairs in the Newman et al., Shields, and Juel-Nielsen studies did not come close to being "reared apart" based upon most people's understanding of this term. This means *there were no valid studies of reared-apart twins to "replicate."* Moreover, these three studies did not use a DZA control group to assess the meaning of their above-zero MZA IQ correlations. For this reason, Segal saw the creation of the MISTRA DZA control group as "an important methodological improvement over past projects." Because the MISTRA MZA and DZA IQ correlations did not differ significantly, we can assume that the earlier studies, had they also used a DZA control group, would have found similar negative results.

With no apparent "space limitations," Bouchard's 2023 article provided an excellent opportunity to finally publish the MISTRA 1990 full-sample DZA IQ correlations to try to prove me wrong. Yet 33 years after the study's publication, Bouchard continued to keep his full-sample control group DZA IQ correlations secret. Why is that?

P-Hacking in the MISTRA IQ Study. A major reason why science is currently embroiled in a replication crisis is *p-hacking*, which is the practice of researchers consciously or unconsciously manipulating definitions and data, either openly or behind the scenes, to transform non-findings into "findings" that reach the conventional .05 level of statistical significance. I showed in my 2023 book *Schizophrenia and Genetics: The*

End of An Illusion that p-hacking, most of it out in the open, was a major aspect of the most frequently cited schizophrenia adoption studies.

The authors of a 2015 analysis wrote that one aspect of p-hacking "occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results." They concluded that p-hacking is "widespread throughout science." Bouchard wrote, "P-hacking can be defined in several different ways and Joseph provides a few. He does not provide, in a full page of text devoted to the topic (588 words), any examples of p-hacking in MISTRA." In fact, in my 2022 article, I gave two specific examples of apparent MISTRA p-hacking:

> "In the area of IQ, the MISTRA researchers appear to have engaged in p-hacking (a) when they failed to publish and assess their control group DZA IQ correlations at the 1990 *Science* study stop point; and (b) when in the same article they selectively reported a method that produced statistically significant results, while failing to report the results of the planned method (MZA-DZA comparison and/or model fitting) which, the evidence suggests, produced statistically nonsignificant results."

The intended genetic natural experiment Bouchard and his colleagues described in 1986, and Segal confirmed in 2012, involved comparing *r*MZA versus *r*DZA and including these results in their model-fitting procedures, as they did in most non-IQ MISTRA studies. Three examples are a 1988 MISTRA personality study, a 1990 MISTRA study of religious interests and attitudes, and a 1991 MISTRA vocational interests study.

Though perhaps not unusual in the academic psychology research culture of that era, where what we now call p-hacking practices were seen by some as minor violations similar to jaywalking, bypassing the DZA group IQ correlations to arrive at desired conclusions was a classic p-hacking maneuver that transformed negative results into positive ones.

Data Collection Stop Point. According to Bouchard, I implied that he violated the rule of "establishing a stated data collection stop point" even though I knew "full well" this "rule" was "not in place when MISTRA was conducted." If the rule was not in place in 1990, it should have been. Establishing a data-collection stop-point rule helps prevent researchers, usually working behind the scenes using their "hidden flexibility," from "peeking" at their data and stopping data collection before reaching the study's planned stop point to achieve desired and publishable results falling below the .05 level of statistical significance. The rule also prevents researchers from collecting data *past* the stop point for the same reason.

Following the publication of the 1990 *Science* article and up to the study's 2000 endpoint, the MISTRA researchers added 25 MZA and 26 DZA twin pairs, for a final total of 81 MZA and 56 DZA pairs. Bouchard continued to withhold the full-sample DZA IQ correlations from publication and prohibited independent review of the raw data. Most likely, he did so because he was waiting for the samples to become large enough to nudge the MZA-DZA comparison under the critical .05 level of statistical significance. At that

point, he could publish the full-sample DZA correlations. However, the evidence suggests that the MZA-DZA comparisons never reached the .05 significance level, which would explain why the full-sample DZA IQ correlations were never published and why Bouchard didn't let Leon Kamin and others "anywhere near" the MISTRA raw data.

Bouchard, in 2023, quoted from his 1998 publication: "The MISTRA IQ correlations have not yet been fully analyzed. We are awaiting completion of the study before conducting a full analysis." Yet he never published this "full analysis" of the IQ data, nor did he claim to have done so.

The Replication Crisis and Academic Psychology. In their well-known 2012 article on research problems in academic psychology, Leslie John and colleagues developed the concept of "questionable research practices," or "QRPs." They found that the "percentage of [psychologist] respondents who have engaged in questionable practices was surprisingly high," and "that some questionable practices may constitute the prevailing research norm."

The MISTRA IQ study p-hacking behaviors I described in my 2022 article appear to match the following four (out of ten) QRPs John and colleagues described in 2012:

- **QRP #1**: "In a paper, failing to report all of a study's dependent measures [the DZA IQ correlations]."
- **QRP #2**: "Deciding whether to collect more data after looking to see whether the results were significant."
- **QRP #6**: "In a paper, selectively reporting studies that 'worked.'"
- **QRP #7**: "Deciding whether to exclude data after looking at the impact of doing so on the results."

The replication crisis has its roots in Bouchard's field of academic psychology, where shoddy, p-hacked studies based on multiple QRPs, blatantly false assumptions, researcher confirmation biases, and financial conflicts of interest were overlooked, promoted, endorsed in textbooks, and even celebrated for decades. Although many researchers did not engage in such practices and produced sound research, in some cases flawed research helped psychologists build their careers, achieve fame, and attain expert status (examples here and here).

We have the case of the famous psychologist and IQ hereditarian Hans Eysenck (1916-1997). The author of a 2020 *Science* article wrote of the discovery that some of Eysenck's publications contained "suspected data manipulation" favorable to the interests of the tobacco industry that funded him. Eysenck was the author or co-author of over 80 books and over 1,000 scientific papers. Prior to his death in 1997, he was the most cited living psychologist and the third most cited psychologist of all time, behind Sigmund Freud and Jean Piaget. As the *Science* article reported, the scandal has "pushed" this "psychology hero off his pedestal."

Previously, the Association for Psychological Science (APS) had given Eysenck its 1994 "William James Fellow Award" in "recognition of a lifetime of distinguished contribution to psychological science." The APS statement said, approvingly, that Eysenck "has allied himself with unpopular positions, such as…the selective contribution of cigarettes to cancer based on personality. …Time and again, the accumulation of facts has vindicated him" (more on the Eysenck case and his retracted articles later).

The American Psychological Foundation (APF, affiliated with the American Psychological Association, or APA) presented Bouchard with its 2014 "Gold Medal Award for Lifetime Achievement in the Science of Psychology." The APF ignored the non-publication of the full-sample DZA IQ data, Bouchard's circumvention of the APA's *Ethical Principles of Psychologists and Code of Conduct* data sharing requirement (see below), and many additional problems I and others have outlined. The APF Award statement described the MISTRA as "groundbreaking and inventive, exciting and controversial," and a "stunning achievement, a body of work in which all psychologists can take pride." In 2018, the APA's Division 14 awarded Bouchard its "Dunnette Prize" for the study of individual differences.

Small wonder that U.S. psychology's research/publication process is in a serious and long-overdue crisis. The field's leaders were more interested in congratulating and awarding each other than looking closely at an awardee's problematic research publications.

Built-In Genetic Bias. P-hacking can also involve technology, including computer programs with built-in genetic biases. In my 2022 article, I showed that in a 2007 MISTRA publication the researchers knowingly, openly, and approvingly analyzed their data using "Mx" software containing such biases, including reducing the statistical weight given to unexpectedly high DZA correlations that didn't fit genetic models. "Genetic confirmation bias was built into the MISTRA computer software program," I wrote. Bouchard responded, accurately, that Mx "is not MISTRA software." That's like being on trial for breaking into someone's car with one's own hammer and seeking acquittal on the grounds that it was actually a neighbor's hammer.

Access to Raw Data Denied

As mentioned, Bouchard has always denied access to critically minded or independent scholars seeking to inspect and analyze the MISTRA raw data, including the DZA IQ correlations and twins' unpublished case histories and information on their degree of separation (see this 1991 letter to *Science* by geneticist Jonathan Beckwith and colleagues). According to the APA's *Ethical Principles*, however,

> "After research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose,

> provided that the confidentiality of the participants can be protected
> and unless legal rights concerning proprietary data preclude their release."

In 2023, Bouchard repeated his long-standing position that the "MISTRA was required by the University of Minnesota Institutional Review Board to gather informed consent from all participants and guarantee confidentiality." Confidentiality, however, can be achieved by anonymization. Providing full-sample group IQ correlations and anonymous twins' raw IQ scores would not violate twins' confidentiality. Jensen himself wrote in 1974 that MZA data "should be published in full….so that quantitative analytical techniques other than those used by the original author can be applied to the data by anyone who wishes."

I argued in my 2022 article that *regardless of the reason for denying access*, when TRA researchers fail to make their unpublished raw data available, we must evaluate their findings with extreme caution, or even reject their findings outright, because TRA studies are extremely difficult to carry out due to changing policies, practices, and social conditions. Most likely, it will never again be possible to collect large enough MZA and DZA samples to conduct a new TRA replication study. Given the study's significant social, educational, and political policy implications, the MISTRA findings could be disregarded based on the researchers' "data-hoarding" practices alone.

Do Reared-Apart MZ Twins Experience Environmental Similarity Leading to IQ Similarity?

Most critics answer yes to this question. In 2023, Bouchard emphatically answered "NO." Nevertheless, in 1985, he and Segal concluded at the end of their detailed "IQ and Environment" review, "As we have found with most other environmental variables, quality of schooling, amount of schooling, and preschool enrichment do have an influence on IQ."

Related to behavioral similarity in general, MZAs are always the same sex, and society socializes males and females to behave differently. For example, same-sex twins will correlate much higher for the behavioral characteristic "lipstick-wearing, yes or no?" than *opposite*-sex twins. "Genes for" lipstick-wearing behavior have nothing to do with it.

Cohort Influences. Even perfectly separated MZA pairs share many *nonfamilial* environmental influences in common. The *cohort effect* concept refers to similarities in age-matched people's IQ scores, behavior, preferences, beliefs, physical condition, and other characteristics caused not by heredity but by experiencing stages of life at the same time, in the same historical period and cultural milieu. In my 2022 article, I presented Table 3 listing 15 shared cohort influences experienced or potentially experienced by MZA pairs separated near birth and first reunited when studied (rare even in reared-apart twin studies). These important influences included socioeconomic status (SES), gender cohort, age cohort, educational methods, oppression/racism/discrimination/privilege, national/regional/ethnic/religious/political culture, and striking physical similarity.

Conflicting Statements on Environmental (Cohort) Influences. Bouchard spent another quarter of his 2023 article addressing each of the 15 influences I listed and argued that, in most cases, research evidence suggests that these influences are "difference-producing, not similarity-producing." As if growing up experiencing several common cultural influences at the same time causes people to behave *less* alike rather than more alike.

We saw that Bouchard and Segal concluded in 1985 that quality and amount of schooling influence IQ scores. Because MZA pairs usually grow up in similar SES environments, and because the quality and amount of schooling vary depending on an adoptive family's SES, it is reasonable to conclude that MZAs' similar SES rearing environments contributed to their IQ resemblance for non-genetic reasons.

Bouchard's 2023 "shared environmental influences are difference-producing" statement conflicts with others he has made since at least the 1990 *Science* article, where he and his colleagues wrote, "The proximal cause of most psychological variance probably involves learning through experience, just as radical environmentalists have always believed." And looking back in 2016, he wrote,

> "Our interpretation of the results of MISTRA was very straightforward. We expected that with regard to psychological traits, monozygotic twins reared apart were similar because their effective environments were similar."

It appears that cohort and other shared environmental influences are similarity-producing after all, "just as radical environmentalists have always believed." (The fact that Bouchard decided to count most environmental influences as genetic influences is irrelevant. It was fallacious to do so, just as it was fallacious when leading behavioral geneticist and SATSA co-investigator Robert Plomin did so in his 2018 book *Blueprint: How DNA Makes Us Who We Are*.)

Overlooked Natural Experiments. TRA researchers and their supporters overlook countless real and potential natural experiments that are difficult to explain on genetic grounds. As one example, the American Amish (population approximately 370,000) are traditionalist Christians known for simple living, plain dress, and a reluctance to adopt many conveniences of modern technology. If pairs of separated-at-birth male MZAs (born at the same time, as are all twins) who grew up in the same Amish community were reunited for the first time at age 40, they would likely display many similarities in personality, IQ, behavior, religious beliefs and practices, sexual behavior, clothing, facial hair, and so on. The reason? Although they grew up in completely different *families*, they were raised in the same behavior-molding *culture* at the same time. For his 2023 argument to hold, Bouchard would have to conclude that growing up in an Amish community would not cause MZA pairs to behave more similarly than if they had been randomly placed into different homes spanning the entire globe.

The "Flynn Effect." I mentioned the much-discussed "Flynn effect" in my article's Table 3. Moral philosopher/IQ researcher James Flynn (1934-2020) showed in the 1980s, at a

time when the MISTRA was well underway, that IQ scores worldwide had been increasing by about three points per decade (0.30 points per year), including supposedly "*g*-loaded, culture-fair" tests such as Raven's Progressive Matrices (increasing 0.50 points per year worldwide). IQ test creators periodically re-norm their tests and make them more difficult in order to maintain a mean of 100 and a "bell-shaped" IQ-score distribution. Flynn documented "massive IQ gains over time [which] revealed that the present generation has a huge IQ advantage over the previous generation." Genetic theories cannot explain these massive IQ gains, but environmental factors such as improved nutrition and healthcare, better teaching methods and increased spending on education, and technological advances can help explain them.

Bouchard responded in 2023 that Flynn believed in "the importance of 'heritability.'" True enough, but Flynn also wrote in a 1987 *Psychological Bulletin* article that his findings indicate that "psychologists should stop saying that IQ tests measure intelligence. They should say that IQ tests measure abstract problem-solving ability..." In any case, the most relevant point is that the MISTRA twins were born at the same time and usually grew up in the same country. Therefore, their educational, learning, and cognitive skills development occurred simultaneously in similar Flynn-effect "huge IQ advantage or disadvantage" environments. Theoretically, due solely to the Flynn effect, if the MISTRA MZA pairs had been born two generations apart, the younger twins would have scored about 15 points higher than their much older yet genetically identical co-twins.

Although Flynn generally accepted twin study heritability estimates, the "massive IQ gains over time" he documented provides an additional non-genetic reason why twins will correlate higher for IQ versus randomly selected pairs of individuals spanning the entire age range. We should add the Flynn effect to the long list of environmental factors that confound genetic interpretations of MZA IQ correlations, further undermining the already extremely shaky MISTRA "MZA correlation directly estimates heritability" assumption.

In his 2009 book *What is Intelligence*, Flynn tried to "solve the paradox of how environment could appear so feeble in the twin studies and yet so potent in IQ gains over time." In my view, the "paradox" Flynn described is easily solved. Massive IQ gains over time are real, and twin study IQ genetic (heritability) findings constitute a century-long scientific illusion that Flynn was unable to recognize. Studies using "classical twin method" reared-*together* MZ-DZ comparisons help sustain this illusion. The twin method depends on the assumption that both types of twins grow up experiencing "equal environments." Critics make a compelling case that this crucial assumption is false, meaning that reared-together MZ-DZ comparisons cannot be interpreted genetically.

Were Age and Sex Influences Controlled For? In 1984, McGue and Bouchard created a formula to correct for age and sex effects on twins' behavioral correlations. "For most psychological, physiological, and medical variables," they wrote, "there are substantial age and sex effects…. Failing to correct for age and sex effects when they exist will result

in overestimation of the twin intraclass correlation." Bouchard thereby recognized that age and sex cohort effects alone constitute "substantial" non-genetic similarity-producing influences on behavior, and that the MISTRA needed to control for these influences. He wrote in 2023, "As Joseph acknowledges, we deal with the issue of sex differences, so that is not an issue." I said they created a "questionable and complicated statistical procedure" to adjust data for age and sex effects. I didn't say their procedure was valid or adequate. Similarity-producing age and sex influences remain an issue.

The Converging Evidence ("Triangulation") Argument

In my article I argued that we must evaluate the MISTRA IQ study based on its soundness and logic, and that it cannot be validated by previous TRA studies or other types of behavioral genetic research. "A psychological study, test, or method," I wrote, "must stand or fall on its own logic and soundness, and cannot be validated by supposed 'converging evidence' from other methods." Bouchard strongly disagreed, and argued at several points that the supposed MISTRA findings "triangulate" with or should be evaluated in the context of other types of behavioral genetic research, including animal research. Examples from his 2023 article are as follows:

> "Many studies in the behavioral sciences use small samples and, consequently, are not 'true experiments,' and these problems bedevil all of them. This problem has been solved in behavior genetics by using replications, multiple corroboration, constructive replications and model fitting."
>
> "No study is perfect, including MISTRA, and that is why research must rely on constructive replication and multiple corroboration (triangulation)."
>
> "Studies of both genetic and environmental influences require a combination of research strategies."
>
> "We included the findings for IQ from the three previous MZA studies….We concluded that 'general intelligence or IQ is strongly affected by genetic factors.' These results were replicated in Sweden [SATSA] two years later…using a design that included both MZ and DZ twins reared apart and together."

Bouchard failed to mention that the SATSA researchers defined their "reared-apart" twins as follows: "By definition, the twins reared apart were separated by the age of 11." About 44% of the SATSA twins were raised by members of the same family, usually with the mother raising one twin, and the mother's sibling or parent raising the other twin. In a 1993 publication, Bouchard saw his own investigation, which assessed and tested twins in person in Minneapolis, as far superior to the SATSA in several respects: "Their instruments are very inferior to ours….Their zygosity diagnosis is entirely by questionnaire and their data collected by mail." By definition, the MISTRA twins had been separated by four years of age, not eleven. Replication means repeating a study's procedures and definitions, matching its assumptions, and observing whether the prior findings are confirmed. Clearly, the MISTRA and SATSA procedures and definitions were different.

"At no point," Bouchard wrote in 2023, "does Joseph refute the converging evidence (e.g., the animal work cited at the beginning of this manuscript) in favor of the hypothesis that genetic factors influence human traits." Bouchard thereby implied that his IQ study's findings don't hold up on their own and depend on findings from other studies, and even other species. Quite an admission.

It wasn't my task nor was it necessary to "refute" the supposed "converging evidence" that I and many other writers have critically examined in publications spanning several decades. My task was to explain in detail why the famous MISTRA IQ study—standing alone—produced no evidence that hereditary factors influence IQ score differences.

I cited the late psychologist Scott Lilienfeld and his colleagues, who observed in 2003 that the "proponents of pseudoscientific claims….typically maintain that scientific claims can be evaluated only within the context of broader claims and therefore cannot be judged in isolation." Bouchard appeared to delight in his gotcha revelation that "citing Lilienfeld is ironic. As a [University of Minnesota] graduate student he gathered psychophysiological measurements from TRAs." If "he were still alive," wrote Bouchard, "he would refute Joseph" on radical environmentalism and pseudoscience.

I already knew that Lilienfeld was an admirer and former student of Bouchard, and a theme of Bouchard's 2023 article was a kind of all-or-nothing evaluation of his study's critics and supporters alike. In Bouchard's eyes, if Kamin, Taylor, Farber, and Joseph were wrong about something, everything else they said must also be wrong and they become disreputable and refuted "pseudoscientists." On the other hand, because Lilienfeld worked in TRA research and admired Bouchard, no aspects of his writings can be used to argue that Bouchard's "multiple corroboration (triangulation)" converging-evidence defense of the MISTRA work is also invoked by discredited pseudosciences in support of their claims. That's not the intellectual world most people live in.

The MISTRA IQ results can be evaluated only through a careful analysis of the study's data, methods, researcher practices, and assumptions, just as claims by the authors of a phrenology study cannot be validated by grouping their study with other methods that supposedly "converge" to predict mental characteristics.

Appeals to Authority

Bouchard began his 2023 article by quoting the renowned scientist Charles Darwin, and ended by quoting Nobel Prize-winning physicist Richard Feynman in support of the idea that my 2022 analysis "is not science—it is, to use Feynman's term, pseudoscience." Let's look at the Feynman passage from a publication that Bouchard called an "astute critique of both physical and social science research," exactly as Bouchard quoted it:

> "The first principle is that you must not fool yourself—and you are the easiest person to fool. So you have to be very careful about that. After you've not fooled yourself, it's easy not to fool other scientists… In summary, the idea is to try to give *all* of the information to help others to judge the value of your contribution;

> not just the information that leads to judgment in one particular direction or another" (emphasis in original).

"The idea is to try to give all *of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another."* Wise words, Professor Feynman! One can assume this Nobel Prize winner would not have approved of Bouchard's decision to keep his unpublished "information" off limits to independent analysts, and to omit the control group DZA IQ correlations from his 1990 IQ study.

Since Bouchard quoted a heavyweight scientist against me, I call on another heavyweight who had doubts that a TRA study is able to disentangle potential genetic and environmental influences on behavior. In a 2007 passage, leading IQ psychologist researcher/author Robert Sternberg wrote:

> "There are methods that can be helpful, such as the method of separated identical twins, but even these methods have their limitations, such as the confounding variable that identical twins tend to be placed in similar, and hence correlated, environments so that effects that may appear to be a result of genetic factors may, in fact, not be a result of such factors."

Wise words, Professor Sternberg!

Conclusion: *Science* Should Retract the MISTRA IQ Study

Leaving aside environmental confounds, lack of proper separation, secret raw data, reliance on false or questionable assumptions, and other problem areas, I argued in my 2022 *Human Development* article that the Minnesota researchers suppressed (omitted and bypassed) their DZA control group IQ correlations to arrive at desired conclusions in favor of substantial IQ heritability (70%). Had they chosen to publish their DZA correlations, they would have arrived at an undesired 0% IQ heritability finding. Nothing Bouchard wrote in his 2023 article leads me to modify this conclusion.

As seen on the Retraction Watch website and elsewhere, academic journals are retracting fraudulent and p-hacked research publications at an increasing rate. Journals have retracted at least 13 articles by Eysenck, the 13th "most eminent psychologist" of the 20th century, and dozens more have been flagged for possible retraction. The journal *Psychological Reports* alone retracted 10 Eysenck publications due to "concerns with the validity of the datasets." *Science* unwittingly published in 2011 a subsequently retracted fraudulent study by psychologist Diederik Stapel, whose retraction count is now up to 58. *Science* reported on the Stapel fraud case in a 2012 article.

While I was writing this article, the President of Stanford University resigned from his post due to charges of research misconduct. As reported in the July 19th, 2023 edition of *Stanford Daily*, "He will also retract or issue lengthy corrections to five widely cited papers for which he was principal author after a Stanford-sponsored investigation found 'manipulation of research data.'" Two of the retracted articles appeared in *Science*. We are entering a new and long-overdue era of increased scrutiny of scientific research publications. The replication crisis has dramatically demonstrated that we cannot rely on

the peer-review process to prevent the publication of methodologically unsound research based on unsupported assumptions and QRPs.

According to 2019 guidelines published by the Committee on Publication Ethics (COPE), "Retraction is a mechanism for correcting the literature and alerting readers to articles that contain such seriously flawed or erroneous content or data that their findings and conclusions cannot be relied upon."

The current *Science* "general policies" for publication, adapted from the COPE guidelines, are found on its website. The following *Science* guidelines attempt to prevent the publication of p-hacked and fraudulent research (all emphasis in original):

> "All data used in the analysis must be available to any researcher for purposes of reproducing or extending the analysis."
>
> "Authors should present results in a complete and transparent fashion so that stated conclusions are backed by appropriate statistical evaluation and limitations of the study are frankly discussed."
>
> *"Rules for stopping data collection*. Did you define rules for stopping data collection in advance (for example, specific intermediary and final endpoints)?"
>
> *"Data inclusion/exclusion criteria*. What criteria did you apply for inclusion and exclusion of data? Were these criteria established prospectively?"
>
> *"Research objectives*. State the objectives of the research, clearly distinguishing pre-specified hypotheses from hypotheses suggested after initiation of the data analyses."
>
> "The *Science* journals generally require all data underlying the results in published papers to be publicly and immediately available. Post-publication embargoes are not permitted, nor are stipulations for readers to contact the authors."

I doubt these *Science* policies were in force in 1990, but whether Bouchard and colleagues played by the rules as they stood then, long before the replication crisis, has no relevance to whether *Science* should now retract the MISTRA IQ study because its authors violated its policies. Regardless of the researchers' intent and the dysfunctional research/publication culture they were required to operate in, there can be no statute of limitations for false-conclusion p-hacked studies carrying huge social, educational, and political implications.

The *Science* policies section addresses the issue of study retraction directly:

> "In cases of identified errors or irreproducibility of research findings reported in a *Science* journal paper, a retraction is likely if the core conclusions are thereby invalidated. An accumulation of errors identified in a paper may cause the

> editors to lose confidence in the integrity of the data presentation, and the paper may be retracted."

Because the "accumulation of errors identified" in the virtually irreproducible 1990 MISTRA IQ article resulted in its "core conclusions" being "invalidated," the Editors of *Science* should retract this article. They should take this step not as a form of punishment or condemnation, but to correct the scientific literature.

In his positive 1976 review of Kamin's *The Science and Politics of I.Q.* (Bouchard mentioned only negative reviews), the late evolutionary biologist Richard Lewontin wrote that Kamin discovered "a pattern of shoddiness, carelessness, miserable experimental design, misreporting, and misrepresentation amounting to a major scandal" in IQ-hereditarian research. And in 1980, Howard Taylor described what he called "The IQ game," by which he meant IQ genetic researchers' "use of assumptions that are implausible as well as arbitrary to arrive at some numerical value for the genetic heritability of human IQ scores on the grounds that no heritability calculations could be made without benefit of such assumptions." Unfortunately, not much has changed since then. IQ hereditarians and others continue to play the IQ game, and critics continue to expose this game now aided by replication-crisis-era terms, concepts, and perspectives.

***

I thank two colleagues for helpful feedback on an earlier draft of this article.

Jay Joseph, Psy.D. is a clinical psychologist practicing in the San Francisco Bay Area. He has analyzed genetic research in the social and behavioral sciences since the late 1990s and is critical of medical models of human psychological distress and dysfunction. He is the author of four books, most recently *Schizophrenia and Genetics: The End of an Illusion* (Routledge, 2023). Many of his online articles can be found at the Mad in America website.