On group differences in the heritability of intelligence: A reply to Giangrande and Turkheimer (2022)

Article *in* Intelligence · March 2023

DOI: 10.1016/j.intell.2023.101737

Bryan J Pesta
Cleveland State University
54 PUBLICATIONS **1,016** CITATIONS

SEE PROFILE

Jan te Nijenhuis
University of Amsterdam
118 PUBLICATIONS **2,792** CITATIONS

SEE PROFILE

Jordan Lasker
20 PUBLICATIONS **69** CITATIONS

SEE PROFILE

Emil O. W. Kirkegaard
Ulster Institute for Social Research
156 PUBLICATIONS **863** CITATIONS

SEE PROFILE

Intelligence 98 (2023) 101737

Contents lists available at ScienceDirect

Intelligence

journal homepage: www.elsevier.com/locate/intell

On group differences in the heritability of intelligence: A reply to Giangrande and Turkheimer (2022)

Bryan J. Pesta [a,*,1], Jan te Nijenhuis [b,1], Jordan Lasker [c], Emil O.W. Kirkegaard [a], John G.R. Fuerst [a]

[a] Independent Researcher
[b] Gwangju Alzheimer's Disease and Related Dementias Cohort Research Center, Republic of Korea
[c] Texas Tech, Lubbock, TX 79409, USA

ARTICLE INFO

Keywords
Scarr-Rowe hypothesis
Intelligence
Heritability
SES
Race differences
Ethnic differences

ABSTRACT

Here we reply to Giangrande and Turkheimer's (2022; G&T) recent critique of a meta-analysis we published in *Intelligence* regarding the Scarr-Rowe Hypothesis and the apparent lack of putative race/ethnic group differences in the heritability of intelligence (Pesta et al., 2020). Our rebuttal is divided into three sections that address ubiquitous misstatements in their critique: Section 1 focuses on conceptual and theoretical points. Section 2 addresses methodological, statistical, and interpretative points. Section 3 provides new analyses suggested by G&T that support our original interpretations. We note that G&T published their critique in *Perspectives on Psychological Science* (PoPS), which did not invite us to respond before their paper was published and our subsequent submission of a rebuttal was not accepted. Our unsuccessful appeal of these events based on possible ethics violations is detailed here (Appendix E). We recognize that this is a controversial area of research with legitimate disagreements and hope our responses maintain a degree of rigor and professionalism that others can emulate.

In 2020, we published a meta-analysis in this journal on the heritability of intelligence across different races and ethnicities (i.e., Pesta, Kirkegaard, te Nijenhuis, Lasker, & Fuerst, 2020). There we found no substantial evidence for the existence of Race/Ethnicity x Heritability interactions. These null effects were contrary to predictions stemming from the Scar-Rowe Hypothesis, at least as we interpreted it. Two years later, Giangrande and Turkheimer (2022; G&T) published an article highly critical of our meta-analysis (and ourselves), together with the editors and reviewers at *Intelligence* who acted on our paper. G&T's critique, however, appeared in "*Perspectives in Psychological Science* (PoPS), rather than in this journal.

Naturally, we initially submitted versions of this rebuttal for publication at PoPS, wherein the Editor (Klaus Fiedler) ultimately desk-rejected us on our second attempt. We appealed the decision and even filed formal ethics complaints with various stakeholders at PoPS, APS, and Sage Publishing. As detailed in Appendix E, neither PoPS nor APS felt we were owed the right to defend ourselves against G&T's scathing critiques, at least not in PoPS (and Sage has yet to come up with a full-fledged response). Instead, our rebuttal finds its home here.

Appendix E details the timeline of events with PoPS, and the main text below focuses on point-by-point rebuttals of G&T's article, organized in three sections. Section 1 focuses on conceptual and theoretical misstatements made by G&T. Section 2 addresses methodological, statistical, and interpretative misstatements made by G&T. Finally, Section 3 provides several new analyses of our original, meta-analytic data. Our goal is to constructively address most, if not all of G&T's substantive concerns.

1. Conceptual and theoretical issues

1.1. Scarr and Rowe's original hypothesis

A major focus of G&T's critique regarded how Pesta et al. (2020) interpreted and used the terms "Scarr-Rowe interaction" and "Scarr-Rowe hypothesis" when conducting their meta-analysis. G&T argued that these terms—coined by Turkheimer and colleagues

[*] Corresponding author.
E-mail address: bpesta22@cs.com (B.J. Pesta).

[1] These authors contributed equally.

https://doi.org/10.1016/j.intell.2023.101737
Received 17 January 2023; Received in revised form 16 February 2023; Accepted 16 February 2023
Available online 8 March 2023

themselves—were only meant to refer to heritability x SES interactions, and not heritability x race interactions. In this section, we highlight serious problems that we see with G&T's interpretation of the rich literature in this area. We show that Scarr, Rowe, and many other important researchers in this domain indeed (and often explicitly) featured race and/or ethnicity when attempting to answer questions about how heritability and "environmental quality" might interact.

Jensen (1968) originally proposed the "threshold hypothesis." It predicts that if certain groups exhibit lower performance on cognitive ability tests, exclusively because of their environment, then they should also exhibit lower heritabilities for cognitive ability. Jensen (1968, *p.* 14) noted that comparing within-group heritabilities provides"one feasible means of directly testing the hypothesis that [Blacks] perform below most other groups on tests of intelligence and scholastic achievement because of environmental rather than genetic differences." Jensen's hypothesis was first tested by Osborne and Gregor (1968), Osborne and Miele (1969), and Vandenberg (1970). Notably, Nichols (1970) also discussed Jensen's threshold hypothesis at length, as his dissertation featured data from the Collaborative Perinatal Project (later analyzed by Turkheimer, Haley, Waldron, d'Onofrio, & Gottesman, 2003). Nichols (1970), citing Vandenberg (1970), noted:

> Since [Blacks] are indeed thought (see Vandenberg, 1970) to be living in a more restrictive environment than whites for the development of intelligence, the heritability of test performance could be lower in [Blacks] than in whites for this reason (*p.* 116).

This quote by Nichols (1970) contrasts starkly with G&T's claims about the relevance of race in this research domain.

Next, Sandra Scarr (writing as Scarr-Salapatek) published her seminal 1971 article, entitled *Race, Social Class, and IQ: Population differences in heritability of IQ scores were found for racial and social class groups.* Scarr credited Arthur Jensen, Steven Vandenberg, and Paul Nichols, among others, for their critical reading of drafts of her paper. Clearly, as the title of her article implies, Scarr was interested in race differences as well as SES differences.

To wit, Scarr's conceptualization of "environmental disadvantage" included nutritional, social, (non-genetic) biological, and emotional disadvantages, not just socioeconomic status. Indeed, Scarr-Salapatek (1971a) cited the reviews of social and environmental disadvantages by Deutsch, Katz, & Jensen (1968), wherein Whiteman and Deutsch (1968) examined 18 social disadvantage variables within the following domains: social background, economic aspects, motivational aspects, family setting, parental interaction, activities with adults, and school experiences. Scarr corroborated her broad interpretation of "environmental disadvantage" by citing pediatric endocrinologist James Tanner. Here Scarr noted:

> In other words, environmental deprivation – in this case nutritional, social, and emotional disadvantages – has … both a depressing and variable effect on the expression of genetic differences among individuals (Scarr-Salapatek, 1973, *p.* 1046).

Regarding race and ethnicity specifically, Scarr additionally proposed a cultural disadvantage (or, more appropriately, "difference") hypothesis (Scarr-Salapatek, 1971a). Scarr noted that "one may question the equivalence of black and white cultural environments in their support for the development of scholastic aptitudes" (Scarr-Salapatek, 1971a, *p.* 1294). Scarr explicitly identified cultural differences as a potential source of environment x heritability interactions on IQ. In subsequent work, Scarr and Barker (1981, *p.* 262–3) argued that race-specific cultural differences existed within social classes:

> Previous research on socioeconomic differences within the two racial groups indicates that SES differences are also an insufficient explanation… The major hypothesis is that black children are culturally less familiar with the kinds of skills and materials required for high performance on typical intelligence tests.

Scarr and Barker (1981, *p.* 263) then pitted a cultural difference model against a genetic one:

> …Major predictions of the generalized cultural-difference hypothesis are…2. The cultural differences of the blacks constitute a "suppressive environment" with respect to the development of the intellectual skills sampled by typical tests, and therefore black children will show less genetic variability in their scores and more environmental variability (Scarr-Salapatek, 1971a)…[and] The proportion of genetic and environmental variability will be the same in both racial groups.

Two things are particularly notable here. First, Scarr and Barker (1981) described racial cultural differences as potentially creating "suppressive environments" for Blacks. Here, Scarr also referred to her prior statements in Scarr-Salapatek (1971a, *p.*1293). Second, Scarr and Barker (1981) proposed that the similarity of heritabilities between groups would be in line with a genetic hypothesis. Specifically, as Allen and Pettigrew (1973, *p.* 1042) summarized, according to Scarr-Salapatek (1971a):

> Either (i) racial differences in intelligence result from environmental disadvantage that simultaneously retards mental development and prevents full expression of genetic differences or (ii) racial differences reflect genetic differences that contribute a similar proportion of variance in all social classes.

We elaborate on this point below, as G&T mentioned it several times in their critique.

This line of comparative heritability research was initially followed by Hodges, Juarez, and Gardner (1976), DeFries et al. (1976), and Osborne (1980). The next major work in this area was published by Van Den Oord and Rowe (1997). These authors analyzed the NLSY children sample (also re-analyzed by Pesta et al., 2020). In addition to SES variables, the authors examined variance component interactions by race/ethnicity. They noted:

> There is ample literature showing that economic and cultural factors may be different in whites compared to minority groups… For this reason we constructed a variable Minority group which was one for Hispanic and black children and zero for white children (*p.* 211).

Race/ethnicity was indeed included as one of their "environmental histories" in addition to family structure and SES.

Next, Rowe, Jacobson, and Van den Oord (1999) examined the effect of parental education on heritability while also comparing racial groups. The authors noted the similarity of heritable influences for Black and White adolescents. Nothing in their introduction or discussion implied heritability moderation would only pertain to SES. Instead, it pertained to environmental quality, where "parental education" could be considered one proxy for it. This study was followed by Guo and Wang (2002), who explicitly dealt with race/ethnicity x heritability interactions.

Next, Guo and Stearns (2002) clearly considered race to be a component of the "social environment" (e.g., Guo & Stearns, 2002, *p.* 897). In fact, Turkheimer himself (Harden, Turkheimer, & Loehlin, 2007) explicitly acknowledged this fact:

> Guo and Stearns…demonstrated that the interaction between parental education and genetic influences on verbal intelligence disappeared if other environmental indices – income, parental employment, absence of a biological father, and race – were included (Harden, Turkheimer & Loehlin, 2007, *p.* 280).

Next, Turkheimer et al. (2003) examined the same sample featured in Nichols' (1970) dissertation. The authors did not report race x heritability results. Instead, they noted that "several previous studies have addressed differential heritability as a function of race, social class, *or* parental education" (*p.* 623) (emphasis added). Moreover, they referenced Scarr-Salapatek (1971a), Scarr and Barker (1981), Van Den Oord and Rowe (1997), and Guo and Wang (2002). They also cited theoretical discussions of environment x heritability interactions in which environmental factors associated with race/ethnicity were explicitly conceptualized as possible sources of heritability modification (e.g., Bronfenbrenner & Ceci, 1994). So, at least 20 years ago, Turkheimer himself believed that race was obviously relevant to empirical studies in this area.

Thereafter, Turkheimer, Harden, D'Onofrio, and Gottesman (2009) coined the term, "Scarr-Rowe interaction," and they did so by citing Scarr-Salapatek (1971a). For context, however, consider the entire passage Scarr wrote here, contrasted with the portion quoted by Turkheimer et al. (2009):

> The environmental disadvantage hypothesis assumes that lower-class whites and most blacks live under suppressive (19, 20) conditions for the development of IQ. In brief, the disadvantage hypothesis states: (i) unspecified environmental factors affect the development of IQ, thereby causing the observed differences in mean IQ levels among children of different social classes and races; (ii) blacks are more often biologically and socially disadvantaged than whites; and (iii) if advantage were equally distributed across social class and racial groups, the social class and racial correlations with IQ would disappear. **The environmental disadvantage hypothesis predicts that IQ scores within advantaged groups will show larger proportions of genetic variance and smaller proportions of environmental variance than IQ scores for disadvantaged groups.** (Emphasis added).

Turkheimer et al. (2009) additionally credited Scarr and Barker (1981), who tested the racial-cultural difference hypothesis, and reported its partial replication. Moreover, Turkheimer et al. (2009) explicitly discussed the interaction Scarr found between race/ethnicity and heritability. Contra G&T's critique of us, Turkheimer et al.'s (2009) characterization of the Scarr-Rowe interaction is undeniably linked to race. To underscore our claim here, note also that Turkheimer et al. (2009, par 38) concluded with a call for research on the interaction between heritability estimates and SES, age, gender, and race:

> Finally, there is also a need to return to the clear theoretical focus that Scarr brought to her early work on this subject in 1971. Now that software is readily available, it would be possible to re-analyze practically every twin analysis that has ever been conducted*, with the familiar variance components moderated by socioeconomic status, or by age or gender or race*. One would not want the field to wind up in the atheoretical tabulation of moderated variance components, without explicit reference to the developmental processes that underlie them. [Italics added.]

Next, the term "Scarr-Rowe hypothesis" (contrasted with "Scarr-Rowe interaction," as described above) was first defined in a pair of articles published in 2015 (two years after we began collecting data for our meta-analysis). Rhea (2015) stated that the Scarr-Rowe hypothesis "posits that for cognitive ability the influence of shared environment increases and genetic influence decreases in adverse environments" (*p.* 598). Consistent with this, Turkheimer, Beam, and Davis (2015) noted that "The Scarr-Rowe hypothesis refers to the possibility that the heritability of cognitive ability is attenuated in relatively poor environments" (*p.* 635). In our read, it is not obvious that the term "poor environments" here referred exclusively to low SES. To wit, consider the term Bronfenbrenner and Ceci (1994) used in contrast to "good environments":

> Accordingly, the differences in developmental outcome (and corresponding levels of h^2) between poor versus good environments are consistently smaller than those associated with low versus high levels of proximal process" (p. 580).

Moreover, Turkheimer et al.'s (2015) claims must be interpreted in the context of those made in Turkheimer et al. (2009), which very clearly linked "Scarr-Rowe interaction" research to studying variance components as potentially moderated by race.

Other authors have referred to the "Scarr-Rowe hypothesis of Gene × Socioeconomic Status (SES) interactions" (Nielsen, 2016; Tucker-Drob and Bates, 2016). Most papers have characterized the "Scarr-Rowe hypothesis" narrowly as the proposed effect of SES on the heritability of cognitive ability. However, others have expanded the hypothesis by using years of education (Baier & Lang, 2019), or school tracking (Uchikoshi & Conley, 2021), instead of cognitive ability as the trait whose heritability is modified. And still others have defined the "Scarr-Rowe hypothesis" more broadly as the effect of adverse environments in general on the heritability of cognitive ability. For example, according to Holden, Haughbrook and Hart (2021, *p.* 5):

> This pattern of moderation aligns with a bioecological model of development (also called the Scarr-Rowe effect), which proposes that supportive environmental contexts (e.g. higher classroom quality) enhance genetic propensity, and that poorer environmental contexts inhibit genetic potential (Bronfenbrenner & Ceci, 1994).

Holden et al. (2021) cite Bronfenbrenner and Ceci (1994), who considered race/ethnicity to be an "environmental context." So this understanding of the Scarr-Rowe effect" is consistent with ours.

In sum, our literature review demonstrates that Jensen, Osborne, Vandenberg, Nichols, Scarr, Hodges, Rowe, Guo, and even Turkheimer himself, were interested in race/ethnicity x heritability interactions (either in-and-of themselves or via SES). Contrast this with G&T, who instead suggested that Pesta et al. "create[d] the false impression that their race- and ethnicity-based analyses are founded on well-established literature" (*p.* 4).

1.2. Research on race/ethnicity and heritability interactions

Another of G&T's central claims was that research in this area has ignored the potential effect of race/ethnicity for decades now, and that "all recent studies on Gene × Environment interaction have examined interactions of heritability and social class, not race or ethnicity" (*p.* 4). However, Turkheimer himself cited Guo and Wang (2002) in 2003, and Guo and Stearns (2002) as recently as 2007. Moreover, contrary to what G&T claimed, recent studies have indeed examined interactions between race/ethnicity and heritability, together with SES interactions (e.g., Rhemtulla & Tucker-Drob, 2012; Schwartz, 2015). For example, Rhemtulla and Tucker-Drob (2012, *p.* 553) noted:

> Because low SES and racial/ethnic minority status covary substantially in the US, we ran an additional model to test whether the gene x SES interaction on mathematics skill could be accounted for by a gene x race/ethnicity interaction. In this model, the main effects of race/ethnicity and SES as well as the effects of their interaction terms with each biometric component were included.

Likewise, a paper coauthored by Turkheimer himself examined interactions between race, SES, and heritability. Tucker-Drob, Rhemtulla, Harden, Turkheimer, and Fask (2011) noted: "Controlling for race and the interaction of race with genetic and environmental factors did not change the pattern of results we found throughout the study" (*p.* 126). Similarly, Halpern-Manners et al. (2020) included "child's race/ethnicity" as an interaction term among others because "studies have shown that these variables can impact estimates of genetic and environmental influences, as well as their relative importance, when modeling results using an adoption design" (*p.* 9).

Given what has been included in Scarr, Rowe, and other investigators' research programs, we do not think it unreasonable for the term "Scarr-Rowe hypothesis" to refer to the totality of what Sandra Scarr investigated. Scarr and others clearly recognized that SES does not fully capture environmental quality. However, to avoid confusion, we will simply refer to these interactions and their proposed cause, first clearly laid out and extensively developed by Scarr-Salapatek (1971a) as 'Scarr and Rowe's Original Hypothesis.'

Understanding how G&T characterized Scarr's research program is useful for resolving many of their misunderstandings about her writing. For example, G&T's introduction noted that:

> …whereas Pesta et al. appear to regard race and ethnicity as reliable indicators of genetic difference, we reject this notion because it fails to acknowledge important sociocultural factors that differentiate racial and ethnic groups" (*p.* 2).

However, Sandra Scarr proposed comparing heritabilities across races because:

> …both [the environmental disadvantage and genetic difference hypotheses] make differential predictions about the proportions of genetic and environmental variance in IQ within lower and higher social class groups" (Scarr-Salapatek, 1971a, *p.* 1286).

Whether racial identities in a particular country track "genetic differences among individuals, ancestral differences among families, or evolutionary differences among populations in ways that allow them to be used as meaningful biological variables" (G&T, *p.* 3) is simply irrelevant to Scarr and Rowe's Original Hypothesis. As for our usage, when we discussed results, we quoted Turkheimer, Harden, and Nisbett's (2017) reference to "socially defined racial groups." To be clear, by "race/ethnicity" we meant socially defined race and/or ethnicity, consistent with the general usage of the term in the wider literature. The degree to which socially and genetically defined groups overlap was and still is wholly irrelevant to the initial research questions we presented in Pesta et al. (2020).

Given G&T's claims about the validity of racial/ethnic comparisons for the heritability of cognitive ability, one might wonder if others have made such comparisons with respect to other traits. The answer appears to be a resounding 'yes.' We have listed several recent studies to this effect in Appendix A. Heritabilities are frequently compared across socially defined racial/ethnic groups. It is unclear why an exception ought to be made when it comes to intelligence research.

1.3. Relevance of race/ethnicity x heritability interactions

G&T also claimed that race x heritability interactions have no bearing on interpreting SES x heritability interactions. This statement is in direct contrast with statements made by Rhemtulla and Tucker-Drob (2012), who worried that SES x heritability interactions might be confounded by race x heritability interactions. Therefore, it is obviously important for modern researchers to test whether SES x heritability interactions are indeed independent of race x heritability interactions. This is not new in the literature, as Scarr-Salapatek (1971a), Turkheimer et al. (2003), as noted by Turkheimer (personal communication, October 4th, 2013), and others have conducted these tests. Also, as discussed above, Pesta et al. (2020) fully acknowledged that certain racial/ethnic groups may be environmentally disadvantaged in biological (e.g., in terms of nutrition, lead exposure, or iodine deficiency) or social ways not captured by SES.

For our meta-analysis, we contacted 25 research teams to request access to their data (See "S2: List of Contacts" in Pesta et al.). The responses were overwhelmingly positive and not a single responding team – some of which included professor Turkheimer – expressed reservations about the research question. Rather, many indicated strong interest in the results.

Another of G&T's key concerns regarded some of Pesta et al.'s conclusions. On page 4, G&T stated:

> The Scarr-Rowe hypothesis, and especially their racialized version of it, has implications for genetic explanations of mean differences in intelligence among racial and ethnic groups. This inference is even less justified than the first… Even if they had somehow provided unassailable evidence against Heritability × Race and Ethnicity interactions, we reject the idea that this would serve as confirmatory evidence for hereditarian hypotheses about mean differences in intelligence.

First, the respective heritabilities of different races/ethnicities have obvious implications for both environmental and genetic explanations of group differences (Jensen, 1998; Warne, 2021) – a point that is clearly congruent with Turkheimer's writings. In fact, in Jensen's (1998) discussion of the relation between within- and between-group variance components, he cited Turkheimer (1990) to this effect.

Second, we did not argue that similar heritabilities across groups "would serve as confirmatory evidence for a hereditarian hypothesis" (G&T, *p.* 4), nor did we suggest that "conclusions about supposed genetic bases of racial and ethnic differences" would "follow logically from findings about Heritability × Race and Ethnicity interaction" (*p.* 4). Pesta et al. (*p.* 11) instead referenced Scarr's reasoning, and noted:

> Scarr-Salapatek (1971) predicted that lower-scoring racial/ethnic groups would have substantially weaker genotype-phenotype correlations (heritabilities) than higher-scoring ones. It was assumed that the environmental factors causing the cognitive disadvantages would attenuate the genotype-phenotype correlations in the disadvantaged groups. The finding of similar genotype-phenotype correlations across groups could be because the alternative genetic hypothesis is correct. Alternatively, the results may imply that: the general model's key prediction is incorrect. Perhaps "environmental disadvantage" between groups does not substantially lower heritability within groups, even when those groups themselves are disadvantaged in a cognitively impactful way. (Pesta et al., 2020, *p.* 11).

We specifically mentioned Scarr's predictions, discussed above, for the environmental disadvantage versus the genetic hypotheses. Given our results, either Scarr's model is correct, and our findings could be interpreted as agreeing with a genetic hypothesis, or Scarr's model is incorrect, and the correct inference is unclear.

Counter to G&T's claims, we merely suggested specifically that environmental disadvantage between groups does not substantially lower heritability within groups as Scarr supposed it would. Obviously, a nature/nurture interpretation of group differences does not "logically follow" from findings of differences or similarities in heritability across race/ethnicity or SES, as there are still viable alternative explanations for this phenomenon. Scarr herself noted the weaknesses of her model (Scarr-Salapatek, 1971b, *p.* 1226). Indeed, empirically questionable assumptions are required such that the causes of differences between groups are a subset of the causes of differences within groups (Scarr-Salapatek, 1971a). Moreover, as Nichols, 1970, *p.* 213–216; see also, Allen & Pettigrew, 1973) noted, adverse environmental conditions could both raise (e.g., more uniform environments due to lack of opportunity) or lower (e.g., suppressive environments) within-group heritabilities. Therefore, we focused our comments on Scarr-Salapatek's (1971a); see also: Scarr-Salapatek, 1973) model and related claims. This is not to say that comparisons of estimated within-group heritabilities cannot *suggest* causes, given ancillary evidence, or cannot be informative, given specific theories. Regardless, inferences about the etiology of group differences based on comparative heritabilities alone are not straightforward.

In sum, Scarr, Rowe, and other authors with significant involvement in the development of this research area clearly intended environment x heritability interactions to go beyond SES and to also involve race x heritability interactions (among other potential forms of interaction). Moreover, these researchers never considered race/ethnicity to be merely a proxy for SES – instead they posited that there may exist additional environmental factors that may be unique to the experience of individuals in different racial and ethnic categories. It was also clear that Scarr and other authors felt strongly that their research designs could offer some information about the causes of racial/ethnic differences in intelligence.

Table 1
Rationale for Pesta et al.'s (2020) racial/ethnic classifications.

Groups	Authors (if atypical)	Our Classification	Rationale and possible social/cultural effect	Relative score
Non-Hispanic Black		Black	Socially defined Blacks represent a different social and cultural group than Whites with generally worse social outcomes. Cognitive disadvantages could be due to effects of discrimination (Guo & Stearns, 2002) and race-related culture (Scarr & Barker, 1981).	Lower
Non-Hispanic Black and Black/White	Scarr, Weinberg, and Waldman (1993)	Black	We followed Scarr and Weinberg (1976) in considering this group to be socially Black. See above.	Lower
Hispanic		Hispanic	Socially defined Hispanics represent a different social and cultural group than Whites and generally have worse social outcomes. Cognitive disadvantages could be due to Latin American language or cultural effects, or biological effects if born in a developing country.	Lower
Spanish Surname	Hodges et al. (1976)	Hispanic	"Persons of Spanish surname" was used in the Census prior to "Hispanic" being formally added by the OMB. See above.	Lower
Non-Hispanic White		White	Reference Group	Reference Group
Asian	Engelhardt, Church, Paige Harden, and Tucker-Drob (2019)	Asian/Other	Asian Americans are a heterogenous group who tend to have better social outcomes than Whites. Social and cognitive advantages could be due to Asian American achievement culture and related social effects (e.g., Lee & Zhou, 2015)	Higher
Mixed or other race/ethnicity (not Hispanic, White, or Black)	Hart, Soden, Johnson, Schatschneider, and Taylor (2013)	Asian/Other	From a cohort that is mostly Asian/Other. Scores above Whites as did Engelhardt et al.'s (2019) group. Overlaps with that above.	Higher
Multiple	Engelhardt et al. (2019)	Multi-racial	No information on specific racial backgrounds. Performed below Whites on IQ tests so not grouped with Asian/Other.	Lower
Non-White	Rhemtulla and Tucker-Drob (2012)	non-White	The groups of all racial / ethnic minorities; on average they have worse social and cognitive outcomes.	Lower

2. Methodological, statistical, and interpretive issues

2.1. Racial and ethnic classifications

Many of G&T's claims were focused on Pesta et al.'s supposed conflation of race and ethnicity. Again, our research was motivated by Scarr's hypothesis that certain groups have lower mean intelligence scores due to poorer-quality environmental conditions that reduce their within-group heritability. It is obviously irrelevant for this purpose if the socially defined groups are ancestrally homogenous. From this perspective, race/ethnicity, just as SES, correlates with poorer-quality environments. The question is, are the potential disadvantages between socially defined races and socially defined ethnic groups completely different? And are they, in turn, completely different from those between social classes? Nothing in our reading of Scarr, nor Rowe, nor other related researchers' work (excepting G&T) suggests as much.

As described in their methods section, Pesta et al.'s (2020) usage of "race" and "ethnicity" involved searching for papers and contacting researchers from around the world to request heritability data for different races and ethnicities. In some countries, like the USA, clear distinctions have been drawn between the categories of race and ethnicity (U.S. Census, 2021). However, in other countries, like the UK, the same groups are sometimes referred to as "races" and sometimes as "ethnicities" (UK Government, 2019). As such, we used the term "race/ethnicity" (the "/" meaning "and/or"). Our use was consistent with those of Rhemtulla and Tucker-Drob (2012), and Halpern-Manners et al. (2020). Our only exceptions, apart from the electronic search descriptions, were in three sentences where we used only "race" (versus "race/ethnicity") in describing data concerning only race. Overall, Pesta et al.'s meaning was clear from the text, and the listed electronic search terms we used. The meaning was also consistent with established usages in behavior genetics research (e.g., Scarr & Barker, 1981; Van Den Oord & Rowe, 1997).

G&T critiqued our use of specific racial/ethnic groupings, in particular the small multiracial and Asian/other samples. However, our method was appropriate, as it is unethical to throw away data when conducting a meta-analysis. Moreover, since there were only partially consistent classifications across samples, we were careful to group logically to minimize relevant heterogeneity, given the available data. The goal was to retain consistency in both social and cultural environments and cognitive performance across studies. Table 1 illustrates Pesta et al.'s classification scheme.

There were sound reasons for grouping "non-Hispanic Black and Black/White" with other Blacks, a classification which often includes multiracial persons, owing to historic hypodescent conventions. This is why Scarr and Weinberg (1976) classified both children with one and two Black parents as socially Black. Likewise, our decision to include Spanish-surnamed individuals in the category "Hispanics" was anything but arbitrary, as "Hispanic" did not become an official U.S. Census term until 1977, when the Office of Management and Budget (OMB) decided to add it. Prior to 1977, the category was, unsurprisingly, "Persons of Spanish surname" (PRB, 2010).

Next came criticisms about how we handled Hart et al.'s (2013) "Mixed or other race/ethnicity" category. G&T devoted considerable space to discussing the classification, as this group could have been included with either the "Multi-racial" or the "Asian/Other" groups. Our actual inclusion decision was based on two facts: (1) in the Florida cohort from which these data came, there were more Asian/Others (Asian, Pacific Islanders, Native Americans, and Not Specified) than multi-racial persons, and (2) this "mixed or other race/ethnicity" group scored on average higher than Whites on measures of cognitive ability. The second point is critically important, given Scarr and Rowe's Original Hypothesis. This group performed better than Whites, like Engelhardt et al.'s (2019) Asian group, but unlike Engelhardt et al.'s (2019) Multiracial groups. Thus, grouping the high-performing "Mixed or other race/ethnicity" group with the high-performing "Asian" group ought to have reduced heterogeneity in the relevant effect, which for Pesta et al. was cognitive test scores.

Relatedly, regarding Asians, Pesta et al. stated (*p.* 7):

> Moreover, given the Scarr-Rowe hypothesis, one would expect a higher value of A for Asians relative to Whites, since Asians score better than Whites on IQ tests in the USA and since they also scored higher in the two samples we had data on. In earlier generations when Whites had higher social status (but still had lower test scores), Asians should have had lower heritabilities.

G&T elided the rationale we provided for the expectation that Asians would have higher heritabilities given Scarr and Rowe's Original Hypothesis: that the Asian samples had higher mean IQs than the White ones. This is implied by Scarr's environmental disadvantage model. It also appears to have been implied by Guo and Wang (2002), when they commented that the non-White heritability was not substantially lower than that of their White sample possibly "because of the large number of Asians in the non-White population" (*p.* 47). The statement about "earlier generations" regarded a different point of ours. It was a qualification of the claim that the opposite prediction might be made for earlier generations of Asians because they were poorer than Whites – and did not pertain to the data we had.

G&T further criticized our "inclusion of multiracial and "non-White samples" that "are racially heterogeneous by definition" (*p.* 7). They apparently overlooked the fact that Van Den Oord and Rowe (1997) examined the effect on heritability of a combined Hispanic and Black racial/ethnic group and that Guo and Wang (2002) included a combined non-White sample in one of their analyses. Pesta et al. (2020), Guo and Wang (2002), and Rowe et al. (1997) were justified in their classifications, given the actual hypotheses that they tested in their respective publications, which also applied to heterogeneous groups such as "non-Whites." We agree with G&T that such groups are nevertheless less than ideal since they are more heterogeneous in cognitive scores and sociocultural circumstances than more homogenous and better-defined groups. But, since there are mean differences in IQ which are often attributed to mean differences in "quality of environment," the hypothesis clearly still applies.

The point above also relates to what G&T considered to be our "most egregious mishandling of racial-group designations" (*p.* 7), namely our use of the small Minnesota Transracial Adoption Study (MTRAS) sample. Specifically, G&T expressed concern about why we grouped Black and mixed-Black/White groups with Blacks instead of an unspecified mixed group. We did this simply because we had specific information about these groups; they were all, according to the original authors, "socially Black," and both the Black and mixed-Black/White groups scored on average below Whites in terms of cognitive ability.

Additionally, G&T took issue with us following Scarr et al. (1993) in computing the heritability in the Black and Black/White adoptive sample. Specifically, Pesta et al. used the correlation between adoptive [White] parents and transracially adopted children. According to G&T, the degree to which transracially adopted children were related to adoptive parents does not count as a measure of genetic influence on transracially adopted children. However, we instead followed Scarr, Weinberg and Waldman (1993, *p.* 553) who concluded that:

> One implication of these heritability estimates is that black/interracial children adopted by white, middle-class families appear to have the same degree of genetic influences on individual differences in their intellectual achievements as do children in the majority populations of the United States and Western Europe.

2.2. Inclusion of samples analyzed by the authors

G&T expressed concerns over Pesta et al.'s inclusion criteria. They critiqued us regarding how we dealt with unpublished studies; our inclusion of a study published in a lower-prestige journal; and re-analyzing original data, such that our re-analysis had apparently not "undergone rigorous peer review" (*p.* 6). As such, it is worth comparing our meta-analysis to Tucker-Drob and Bates' (2016), who examined SES x heritability interactions, and whom G&T cited *uncritically*.

Tucker-Drob and Bates analyzed 14 samples. Counts for their classifications of "Results were reported in the original article," "Data were reanalyzed by the original authors," "Data reported in the article were reanalyzed by the current authors," and "Raw data were analyzed by the current authors" were 1, 6, 2, and 5 paper(s), respectively. Accordingly, results from just one of their samples had undergone what G&T called "rigorous peer review" (i.e., "the results were reported in the original article"). Moreover, results for seven of 14 samples were based on Tucker-Drob and Bates' (2016) own re-analyses.

For Pesta et al.'s meta-analysis, the counts were: 6, 8, 1, and 1 paper (s), respectively. Results for only two out of 16 of the samples were based on re-analyses. Thus, our meta-analysis was based on fewer unpublished studies than that Tucker-Drob and Bates (2016) conducted.

In manuals for systematic reviews and meta-analyses, researchers are encouraged to put significant time into searching for unpublished studies due to publication bias. For example, in their *Handbook of Meta-analysis*, Schmid, Stijnen and White (2020, *p.* 2) noted:

> A systematic review encompasses a structured search of the literature in order to combine information across studies using a defined protocol to answer a focused research question. The process seeks to find and use all available evidence, both published and unpublished, evaluate it carefully and summarize it objectively to reach defensible recommendations.

Indeed, the percentage of unpublished studies in a meta-analysis is an interesting indication of how hard the researchers worked to find all the research ever carried out on a given research topic.

G&T suggested that including more data in the meta-analysis reduced statistical power: "Other factors reduce power as well: Sample sizes in some of the included studies were extremely small (e.g., only ten twin pairs were included in the Black sample from Woodley of Menie et al., 2015; Table 2)." However, the inclusion of these ten twin pairs was proper, as throwing them out would fundamentally violate systematic review and meta-analysis' goal of collecting as much of the existing data as possible. Moreover, while the inclusion of a sample could reduce random effects power via increasing I^2, removing that study did not change our result: I^2 was 0 (0 – 0.355) before, and 0 (0 – 0.362) after removal.

2.3. Heterogeneous measures of cognitive ability

Like us, Tucker-Drob and Bates (2016) included a mix of achievement and cognitive tests in their analyses. We adopted Tucker-Drob and Bates' (2016) method of including "achievement or knowledge test vs. intelligence test" as a moderator. Cook and Campbell (1979), in their classic and highly influential book, stated that researchers should strive for various operationalizations of their constructs. Therefore, having heterogeneity of cognitive ability measures might be desirable.

A common issue in meta-analyses is that multiple effects will be reported for the same sample in a study (Moeyaert et al., 2017). There is ongoing discussion about how to handle multiple effects within studies (Song, Peacor, Osenberg, & Bence, 2020). In contrast with Tucker-Drob and Bates (2016), Pesta et al. dealt with this by averaging estimates within samples before meta-analysis, which is the traditional 'simple' approach. This methodological choice was theoretically grounded and well-documented. We chose this method based on the research questions, the number of samples involved, the characteristics of the studies, and the availability of data (e.g., lack of covariance matrices for the outcomes, let alone the outcome ACE estimates), the limitations of available statistical packages, etc. An alternative is to perform a multilevel meta-analysis, but our method has the effect of decreasing heterogeneity within samples by assigning each independent sample one estimate. We took a conservative approach in the sense that standard errors were overestimated, meaning that precision was necessarily underestimated. Moeyaert et al. (2017, *p.* 12) noted:

> Averaging the effect sizes and the sampling variances within a study is probably the most 'simple' method to handle dependent effect sizes within a study. However, this study indicates that there are definitely better alternatives to use, because the AV approach is too conservative. When using AV, standard errors are in general overestimated…

However, our focus was on variance component means, for which the averaging method produces unbiased estimates (Moeyaert et al., 2017; Song et al., 2020). We did not test if means were statistically different for which unbiased standard errors would be necessary. G&T further pointed out that this method precluded examination of within-sample heterogeneity. Not testing Scarr's hypothesis *enough* is a potentially disingenuous criticism given the thrust of their argument, that our research project was fundamentally misguided. That said, we agree that this would be an interesting topic for future research when more data are collected; therefore, we provided a table with study data (Table S20b) in Pesta et al. for this purpose.

Other points: (1) Contrary to G&T's claims regarding MTRAS, Pesta et al. did not base variance components on educational level. Instead, we used the adoptive mid-parent WAIS-R or WISC-R scores. (2) Our handling of the Scarr-Salapatek (1971a) estimates was based on logical, well thought out decisions given Scarr-Salapatek's (1971a, p. 1268) "radical decision." (3) As for Mollon et al. (2021) and as clearly noted in the text, these were g-factor scores extracted from 14 tests, not an average of subtest scores.

2.4. Meta-analysis issues

G&T wrote several pages highly critical of our meta-analysis. They argued that we included "irrelevant, unsuitable, or low-quality studies" (*p*. 9). Since they did not specify which studies should be thrown out of the meta-analysis, we could not strongly quantify the influence of the problems they raised; we can only address the point more broadly.

All meta-analyses are open to the criticism that original studies included in them are suboptimal because, for example, some of them were carried out decades ago (Hunter & Schmidt, 2004; Schmidt & Hunter, 2015). An alternative would have been to limit meta-analysis to studies supplying all information required, but that would have meant a massive reduction in the amount of data capable of being aggregated. Consensus is to include studies despite missing information so a larger meta-analytical database can be built, allowing stronger conclusions compared to a meta-analysis that only includes a handful of studies. By building a larger database, unsystematic biases may cancel themselves out.

2.5. Statistical issues

2.5.1. Constraining ACE estimates may bias results

As G&T noted, Pesta et al. adopted the common practice of constraining ACE estimates to fall between the theoretical bounds of 0 and 1. In contrast, G&T cited a recent paper advising against this practice in the context of Scarr-Rowe effect investigations. However, Pesta et al. had reasonable theoretical and practical reasons for using this method. Theoretically speaking, it makes sense to include negative heritability estimates if heritability in a sample is truly negative. We have strong reason to believe it is not. The negative estimates found were very likely due to sampling error since they occurred only in small samples with accordingly low statistical power. As such, it may not make sense to include negative estimates, and "truncation at zero (or rejection of negative estimates) is warranted and guarantees improved estimates in, say, mean squared error" (Steinsaltz, Dahl, & Wachter, 2020, *p*. 346). From a practical point of view, many of the published results and many of the results provided by the original authors were already constrained between 0 and 1. It was typically impossible to remove constraints from these results given the provided data. Thus, for methodological consistency we constrained the other values.

Regarding the actual data, G&T stated (*p*. 9):

> For the additional four studies (in addition to the five discussed above), Pesta et al. dismissed the negative ACE estimates, somewhat cryptically, as not being "an issue" because "all unstandardized values [were] between 0 and 100" (supplementary Table S5b).

In Supplementary Table S2 of Pesta et al. (2020), we listed samples where it was "possible to find/use ACE estimates with negatives" because the authors did not constrain estimates between 0 and 1. These were 9/16 samples (i.e., for the seven other samples, the ranges were constrained from 0 to 1). But these nine were samples for which it was *possible* to examine if there were negative ACE values — not for which negative values actually existed, which is what G&T claimed.

Of these nine samples, four had ACE estimates that were all nonnegative (i.e., "All unstandardized values were between 0 and 100"). This left four remaining samples (or five if including the Scarr and Barker (1981) sample on account of negative values for some subtests in the White sample), all of them small. For example, consider Pesta et al.'s "Black" group. The harmonic N for the five samples with some negative ACE values (for one or the other group) compared to the total Black harmonic N was 895/13,977 or only 6.4%. For Whites the equivalent number was just 1788/26,393 = 6.8%.

The net effect of excluding negative values was trivial, as we noted. For the record, here we provide estimates with and without negatives included in Appendix B, along with the meta-analytic means for these five samples. For Whites, there was no difference. For Blacks, the heritability was *higher* when negatives were included. However, given that these samples constituted only 6–7% of the total kinship samples, the difference in this subset of samples did not influence the conclusions we drew here. Note, also, that the formula for standard errors uses the raw MZ and DZ correlations, not the standardized ACE estimates, so the estimated standard errors were unaffected. Thus, given the entirety of the data available and the points noted above, our considered decision on how to handle negative variance component estimates was logical, justifiable, and ultimately unimportant.

2.5.2. Imputation of missing means, variances, and effect sizes

G&T criticized our mean calculations, suggesting our method "raises several red flags" (*p*. 9). We only used the means in our secondary analysis, yet G&T described the results of this analysis as "one strong finding," which "stands out," was "striking," and was "by far the most robust reported in the article" (*p*. 10), providing "unequivocal evidence for the hypothesis the article presumes to be denying" (*p*. 12). Logically, the results of these "additional" analyses could be no more reliable than the estimates of the heritabilities and means they were based on, which G&T appeared to roundly reject.

G&T also criticized our weighting procedures. However, as Pesta et al. noted, we used two alternative weights, which nonetheless produced equivalent results:

> Because our estimated standard errors were imprecise, we also tried weighting by the harmonic Ns of the samples as an alternative. Using these alternative weights did not alter our results (*p*. 4).

Notably, the same point made above applies here since in conducting the additional analysis, which G&T considered "by far the most robust," we weighted values by the Satterthwaite approximation of the pooled error for heritability.

G&T (*p*. 7) noted further:

> Although it is not clear that these percentages make sense, taken at face value they suggest that >100% of the variance in meta-analytic estimates was attributable to sampling error. Obviously, this is absurd, and such patently unreliable estimates cannot make a systematic contribution to the meta-analysis… Again, the percentages of variance explained by sampling error for the differences in heritability were absurdly large (3033 and 1414 for the multiracial and non-White samples, respectively).

We carried out a Schmidt and Hunter-style meta-analysis. This technique has been used in hundreds of meta-analyses, and most prominently, in Industrial and Organizational Psychology and many highly cited and influential papers. When the number of data points is limited in a meta-analysis, it is frequently found that >100% of the variance between the data points is explained by sampling error. This phenomenon has been known for decades and is called 'second-order sampling error' (Hunter & Schmidt, 1990). If the number of data points is too small to make highly reliable estimates of the amount of variance explained, the standard interpretation is that 100% of the variance is explained by sampling error (meaning there is no variance left for moderators). Hunter and Schmidt (1990) suggested that when second-order sampling error occurs, researchers should try to increase the number of data points to run an updated meta-analysis.

Finally, G&T claimed that variance due to sampling error was high at, respectively, 4.38, 34.45, and 33.22% of the variance between the data points. These actually correspond to relatively low percentages. Sample size does not do a good job explaining the variation between data points in this meta-analysis.

G&T's apparent lack of familiarity with how Schmidt and Hunter-style meta-analysis is conducted shows up again in their suggestion that Pesta et al. mischaracterized the resulting findings for the multiracial and non-White groups. In meta-analysis, it is common to start with an overall analysis of all the data points, by computing the weighted mean of the data points and the variability of the data points. Often a clear picture of the data emerges at this first step in the sense that sampling error explains the variability in the outcomes, so there is little more than curiosity to motivate splitting the database into meaningful groups.

However, it is also possible that an unclear picture emerges with large differences between data points not explained by sampling error. In that case, the meta-analytic database is often broken up into groups. The weighted means are computed within groups and the variability of the data points within groups is then computed. This might (in some cases) result in a meaningful group represented by only one data point. Thus, G&T's suggestion that Multiracial and Non-White groups consisted of only one data point and their results therefore did not constitute outcomes from a meta-analysis evinces misunderstanding of Schmidt and Hunter's style of meta-analysis.

2.6. *Interpretations*

2.6.1. *The expected differences given Scarr and Rowe's original hypothesis*

G&T next criticized Pesta et al.'s interpretation of their results. We did not have a clear quantitative statement on the expected heritability differences to compare our results to because Scarr-Salapatek (1971a) never formulated one. Scarr only provided an illustration (Fig. 1 of Scarr-Salapatek, 1971a) and a comparison with animals raised in normal versus enriched environments ("The percentage of genetic variance in the scores of standard-cage-reared animals was one-fourth that of animals with enriched environments (10 percent versus 40 percent)" (*pp.* 1293–4) alongside various qualitative discussion. For the latter, Scarr-Salapatek noted that: "genetic variability is important in advantaged groups, but much less important in the disadvantaged… the proportion of genetic variance in the aptitude scores of black children is considerably less than that of the white children, as predicted by model 1" (*p.* 1294).

Because Scarr did not provide a clear quantitative statement about expected differences, Pesta et al. wrote:

> One could consider these differences in light of the effect originally reported by Turkheimer et al. (2003). When treating SES as a dichotomous variable, Turkheimer et al. (2003) reported that "the low SES group had a h^2 of 0.10 while the high SES group had a h^2 of 0.72. This represents a large effect by conventional standards. We find nothing like this in the present meta-analysis… Alternatively, one could compare the effects here to those that exist between age groups. Plomin et al. (2014, *p.* 202) reported that heritability increases significantly from approximately 40% in childhood to 80% in late adulthood. This ΔA represents a medium-sized effect, which we do not see here in the context of differences between self-identified racial/ethnic groups. (*p.* 7).

Since no one has provided a specific quantitative prediction for the magnitude of heritability differences, we cannot statistically compare results with the preexisting predictions of others. This lack of a benchmark is why we called for modeling expected heritability differences:

> Proponents of the Scarr-Rowe hypothesis should try to model their predicted effects regarding group differences more explicitly" (p. 11). The point was that perhaps expected differences were smaller than we could detect — or, since Black and Hispanic heritabilities were higher than White ones, that expected differences were smaller than we could rule out. In the absence of clear quantitative predictions, we can only say that claims about "much less," "considerably less," or "markedly lower heritabilities," are not supported, and that substantial race/ethnicity x heritability interactions "likely do not exist."

On the above point, we note that G&T did not provide any statistical tests to back up their claim that it "appears probable" that our analysis was too underpowered to detect differences. Obviously, what is relevant are differences of the size predicted by Scarr and Rowe's Original Hypothesis (i.e., effect sizes, versus merely statistically significant differences, which are simply a function of sample size).

G&T (*p.* 11) next claimed that we dismissed contradictory results:

> First, Pesta et al. cautioned that "results [of the additional analysis] may not be robust, owing to possible confounding factors between samples (e.g., age differences)" (p. 7). Pesta et al. did not mention such confounds when discussing their other meta-analytic results. In contrast, they asserted that "the design here is strong, as the groups are matched on several background variables" (p. 7) and reported no evidence of moderation by demographic variables.

Their claim here is wholly inaccurate. "The design here is strong" referred to our "meta-analysis of matched groups" in contrast to our "meta-analysis of unmatched groups." The meta-analysis of matched groups represented the difference (denoted: Δrho^2) between the meta-analytic means for those samples for which we had matching data, thus limiting between-sample variability. This meta-analysis was contrasted with our meta-analysis of unmatched groups, which was the weighted average of heritability estimates for each race/ethnicity across all samples. Contrary to what G&T claimed, Pesta et al. did mention "such confounds" when discussing our unmatched meta-analytic results. Specifically, we noted that these meta-analytic estimates were "not directly comparable" with each other. We said: "We analyzed all samples in Table 3. The estimates, however, are not directly comparable because the groups differed in terms of the samples in which they participated" (*p.* 7).

We provided a parallel but weaker caution with the heritability × group difference results based on within-sample versus between-sample analyses, noting that the latter "may not be robust." The heritability × group difference analysis across all samples paralleled the unmatched ACE analyses. After all, we were correlating within-sample differences in heritability and cognitive scores across samples which differed in age, methods of h^2 estimation, cognitive measures, and the sets of races/ethnicities with data, etc.

2.6.2. *Heritability × group difference interaction*

Contrary to what G&T claimed, we did not dismiss the results of our novel group difference heritability analysis. Rather, in Pesta et al.'s discussion, we concluded that the heritabilities among Whites, Blacks, and Hispanics were similar. This coincided with the reported heritability × group difference interaction. The only sentence that might suggest otherwise was in our abstract, where we said: "We found that White, Black, and Hispanic heritabilities were consistently moderate to high, and that these heritabilities did not differ across groups. At least in the

United States, Race / ethnicity × Heritability interactions likely do not exist" (*p.* 1). But "Race / ethnicity × Heritability interaction" in this passage clearly referred to the main effect of race/ethnicity on heritability. The idea of a heritability × group difference interaction was our own auxiliary hypothesis derived from our close reading of Scarr's and others' work.

Regarding heritability × group difference interactions, we justifiably cautioned that "we found evidence consistent with the interaction… however several confounds existed, such as differences in age, methods of estimated h^2, and differences in cognitive measures" (*p.* 10). In Appendix C, we show the effect of controls for these three variables. While, in the combined sample, the relation between Δh^2 and d remain significant, we would not characterize these results, let alone those in Pesta et al., as providing "unequivocal" evidence in support of a modified version of Scarr and Rowe's Original Hypothesis. This is because these analyses control for only a few factors that differ between samples and also because the number of independent pairwise comparisons is small. Rather, we reiterate our statement above. Adding to the ambiguity, Pesta et al. did not find evidence of a heritability × group difference interaction when looking at subtest scores within samples. While G&T criticized this analysis, arguing that "the Scarr-Rowe hypothesis makes no particular predictions about subtest effects" (*p.* 11), the idea for these analyses was based on Scarr and Barker (1981, *pp.* 283–5).

Is it possible there could be a heritability × group difference interaction despite no race/ethnicity × heritability interaction? This interaction, in a way, would be consistent with Scarr's model. But that is what Pesta et al. indicated by "this pattern of correlations may represent a Scarr-Rowe effect of sorts" (*p.* 8), and "a caveat is in order…" (*p.* 10). To be clear, though, we are not discounting our heritability × group difference analyses. We think this is a novel alternative way to test for Environment x Heritability interactions.

One obvious reading of G&T's critique is that the strongest analysis would involve computing the meta-analytic means for Δh^2. By virtue of meta-analyzing difference scores for subgroup comparisons, groups are matched on factors that varied between samples. This is a stronger test of Scarr and Rowe's Original Hypothesis than the heritability × group difference analyses, since it avoids any potential problems ("red flags," in G&T's words) with imputed means and standard deviations for the cognitive differences and since it does not involve correlating across groups or the use of unmatched groups. A disadvantage of difference scores is that "calculation of a change score requires measurement of the outcome twice and in practice may be less efficient for outcomes that are unstable or difficult to measure precisely" (Higgins et al., 2019, *p.* 252). Pesta et al. (2020) did not meta-analyze difference scores for this reason: it was judged that the individual h^2 estimates were imprecise and unstable.

However, G&T clearly disagreed, since they interpreted our heritability × group difference analyses, which were based on difference scores, as strong and a source of robust evidence in support of their favored hypothesis. For this reason, we computed the meta-analytic means for Δh^2 for the White and Black, the White and Hispanic, and the Hispanic and Black groups. They were weighted by the inverse of the standard error of heritability. In parentheses, we have also reported the unweighted results. The meta-analytic mean heritability differences were White minus Black: 0.02 (−0.03); White minus Hispanic: −0.02 (0.00); Hispanic minus Black: 0.05 (0.05). These are trivial to small differences. So, to repeat what was said in Pesta et al.: while the heritability × group difference analyses may suggest an effect consistent, in a way, with Scarr and Rowe's Original Hypothesis, the data do not show the kind of effect Scarr predicted.

2.6.3. *Interpretation of other results*

G&T stated that "Pesta et al. … concluded that intelligence differences are not driven by environmental factors" (*p.* 11) and speculated that "the between-groups differences in intelligence that Pesta et al. reported may still reflect the effects of environmental disparities" (*p.* 11). Again, we did not make any claims about environmental or genetic causes. Rather, we made a claim about heritability differences and Scarr-Salapatek's (1971a) claim that equal heritabilities would be inconsistent/consistent with an environmental disadvantage/genetic hypothesis. Therefore, we were interested in the score differences when "the point in the regression plots where heritability was equal between higher- and lower-scoring races/ethnicities" (*p.* 10). Scarr's model could be wrong though, as we suggested it was.

Finally, G&T claimed that "given the extensive problems we have described and evidence that directly refutes their hypothesis, one might expect that Pesta et al. would exercise some caution when summarizing their findings…" (*p.* 11). This statement should be contrasted with what we actually said (*pp.* 10–11):

> [1] Ethnic groups did not substantially differ in the heritability of intelligence… [2] A caveat is in order regarding whether the Scarr-Rowe interaction actually exists. When we looked across samples, we found evidence consistent with the interaction…" [3] "Regardless, we conclude that ACE × SES interactions, when found, are not being driven by ACE × Race / ethnicity interactions. [4] Our general findings are at odds with the predictions of Scarr-Salapatek's (1971a) environmental disadvantage hypothesis." [5] "The finding of similar genotype-phenotype correlations across groups could be because the alternative genetic hypothesis is correct. Alternatively, the results may imply that [Scarr-Salapatek's, 1971a] general model's key prediction is incorrect. Perhaps "environmental disadvantage" between groups does not substantially lower heritability within groups. [6] "Our meta-analysis reveals that the heritability of cognitive ability is generally moderate to high for Whites, Blacks, and Hispanics in the United States." [7] "We also found that differences in heritability across these three groups were mostly trivial." [8] "Nonetheless, we cannot rule out the existence of modest differences in population parameters in our analyses."

Readers can verify that all eight statements are correct. The only difference is that we described the higher heritability of Hispanic cognitive scores relative to Whites as trivial. But, in doing so, we downplayed results that contradicted Scarr's hypothesis.

To be clear, some of the present authors concur – in line with a solid majority of surveyed mainstream intelligence researchers (Rindermann, Becker, & Coyle, 2020) – that American racial and ethnic group differences in intelligence have a nonzero genetic basis. The point, though, is that we did not argue this based on differential within-group heritabilities. Rather, we noted that such a conclusion presumes, as Scarr explicitly did, that environmental disadvantage between groups substantially lowers the heritability of cognitive ability. We doubt this premise is entirely correct. Pesta et al. nevertheless pointed out in both their introduction and conclusion, that knowledge of race/ethnicity-specific heritabilities is important for several reasons. Here, we gave an example of interpreting polygenic score validities.

3. Novel analyses

3.1. *Outlier and moderator analyses*

3.1.1. *Method*

G&T argued many of the data points in the Pesta et al. meta-analytic database were invalid because they were based on studies that were methodologically flawed in one or more ways. We already considered their arguments in detail, so it is unnecessary to reanalyze the data accounting for G&T's criticism. However, we empirically tested whether it was plausible that there was an abundance of methodological flaws in the studies used in the Pesta et al. meta-analytic database.

Our reasoning was as follows. Sound studies yield correct data points, but studies with methodological flaws yield incorrect data points. So, a comparison between the studies with the most noteworthy methodological flaws and the others should clearly show differences in outcomes. Then, the question is how large these differences should be, and a close reading of G&T provided the answer. When comparing the flawed studies to the sound studies, one would not expect the flawed studies to yield scarcely different outcomes.

We conducted statistical analyses to test this hypothesis for outliers and moderators. First, when G&T focused on one data point, we ran an outlier analysis because running a moderator analysis on only one data point yields only limited information. Second, when G&T focused on two or more data points, we ran a moderator analysis and analyzed inverted funnel plots.

The Hunter and Schmidt meta-analytic program produced the mean sample size-weighted observed correlation and the *SD* value of the observed correlations, which allowed meta-analytic means to be recomputed based on the removal of the "questionable" data points – according to G&T. These could then be used to check for discrepancies (in terms of differences expressed in *SD* units) via comparison with the original value. Outliers were defined as those 4 and 7 *SD*s above or below the mean, strong outliers were defined as being >7 *SD*s and up to 10 *SD*s above or below the mean, and extreme outliers were defined as being >10 *SD*s above or below the mean. It should be noted that having strong/extreme outliers in the meta-analytic database strongly influences the size of the *SD*, so when a visual inspection of the meta-analytic data points suggested a strong or extreme outlier, we recomputed the *SD* after leaving out the suspected strong or extreme outlier and computed the distance from the mean using the new *SD*.

Pesta et al. reported inverted funnel plots, with effect sizes on the x-axis and precision on the y-axis for use in their publication bias analyses. Data points on the far left and the far right of the distribution and within the inverted funnel have small sample sizes and come with a large amount of sampling error. When outliers, strong outliers, and extreme outliers were found, we inspected the inverted funnel plots from Pesta et al. to see whether the data point was still inside the inverted funnel. Being inside strongly suggested a large amount of sampling error was the cause of the data point's position in the distribution.

A moderator in the Hunter and Schmidt tradition yielded substantial differences in means between moderator categories and substantially increased the percentage of variance explained between the data points in the moderator categories compared to the situation without a tested moderator. We used the meta-analytic tables presented in Pesta et al. and added information on the values of moderators. Two of the present paper authors independently rated the values of moderators after a close reading of G&T and then compared their outcomes; they discussed this until consensus was reached. As in Pesta et al., we used the Schmidt and Le (2004) meta-analytic software. The outliers and moderators were as follows:

(1) The Minnesota Transracial Adoption Study.

G&T wrote that Pesta et al.'s data point from the Minnesota Transracial Adoption Study suffered from various methodological flaws. A data point that was mishandled should logically have outcomes that are markedly different from the rest of the data points, so we tested whether this data point was an outlier.

(2) Inclusion of data from unpublished studies.

G&T wrote: "Unpublished work is commonly included in meta-analyses. However, Pesta et al.'s disproportionate inclusion of their own, previously unpublished and insufficiently reviewed results in their own meta-analysis jeopardizes the trustworthiness of their meta-analytic findings. Of the 16 independent samples included in the meta-analysis, two were based on previously unpublished studies by the authors of Pesta et al. (2020)."

We created two categories: one of the data points from studies published in regular journals and the second of data points from unpublished studies. The latter category contained two data points: Fuerst (2014), an unpublished analysis, and Fuerst and Dalliard (2014), which was published in a lower-impact journal, which G&T incorrectly described as 'self-published'. However, both used data from longitudinal projects, and Pesta et al. used the most recent data from the projects, so it would arguably be best to describe both data points as based on unpublished studies.

(3) Negative ACE estimates.

To address this issue, we created two categories: one of data points based on ACE values that did not originally have negative estimates and ones that originally had, or were based on subtests that had, negative values. The latter category contained three data points for Whites and four data points for Blacks.

(4) Averaging within samples.

Multiple estimates per sample are a common problem in meta-analysis. We took the traditional solution by first averaging estimates within samples. G&T criticized this methodological choice. To address them, we created two categories: one of data points based on ACE values that did not involve averaging within samples and one that involved averaging within samples. The latter category contained nine data points.

(5) Imprecise Group Designations.

G&T stated that Pesta et al. made a mistake by "… using group designations that are imprecise at best and incorrect at worst." We previously discussed G&T's problematic interpretation of Scarr and Rowe's original hypothesis, specifically that it required the use of genetically defined racial/ethnic groups instead of socially defined ones. Nonetheless, we could test their claim, given that interpretation. The White groups were most likely close to 100% European and other West Eurasian, so the category White was probably ancestrally precise. U.S. Blacks have on average 75–85% Sub-Saharan African ancestry and 15–25% European ancestry, so the Black category was likely modestly precise as a proxy for continental race and very precise as a proxy for the African American ethnic classification at large. The most ancestrally imprecise group designations were Hispanics, Asian, Asian/Other, Multi-racial, and Non-White. The Asian groups could be Northeast Asians, South Asians, or even Central Asian, so their samples constituted a potentially highly ancestrally diverse category. For our moderator analyses, we used three categories:

a) Precise group designation: Whites (16 samples).

b) Moderately precise group designation: Blacks (excluding MTRAS sample) (14 samples).

c) Imprecise group designation: Hispanics, Asian, Asian/Other with Multiracial, Multi-racial, Non-White, MTRAS (12 samples).

3.1.2. *Results of the outlier and moderator analyses*

With an outlier defined as being at least 4 *SD* away from the meta-analytic mean, Table 2 shows that the Black/interracial data point from the MTRAS is at between 0.26 *SD* to 0.81 *SD* from that group's various means, so there is insufficient evidence it was an outlier. All three estimates were close to the meta-analytic mean. Thus, G&T's hypotheses were unsupported.

We next split the collection of data points in two and carried out separate meta-analyses. We then checked whether there were substantial differences between the mean sample-size weighted means of the two groups and whether the amount of variance explained by sampling error increased in all categories. Additionally, we checked the direction and magnitude of heritability differences for the two moderator categories because G&T implied that our analytic decisions worked against finding a race/ethnicity x heritability interaction. If G&T were correct, their preferred moderator categories (i.e., not averaged within sample) should show more of a race/ethnicity x heritability interaction than the alternative categories (i.e., averaged within sample). To make this comparison, we calculated the Black and White Δh^2 for each of the moderator categories (i.e., not averaged within samples and averaged within samples) and then subtracted the Δh^2 for G&T's preferred category from the Δh^2 for the category G&T criticized. This yielded the relative Δh^2. A positive value indicated that G&T's preferred category showed more of a race/ethnicity x heritability interaction. Detailed information about this is found in Table 3.

The Cochrane Handbook for meta-analysis (Higgins et al., 2019) stated that there should be at least ten data points for a meta-analysis and ten data points for a moderator analysis. There were 15 data points for Blacks which could be matched by 15 data points for Whites, 15 data points for groups that we classified as having an ancestrally precise group designation, 12 data points for groups that we classified as having an ancestrally imprecise group designation, but there were only seven data points for Hispanics, so we did not run moderator analyses for this group.

Testing the moderator precision of group designation included an analysis where we compared data points with an ancestrally precise group designation to data points with an ancestrally imprecise group designation, and we carefully matched the data points with an ancestrally imprecise designation to the data points with an ancestrally precise group designation. However, the Hart et al. (2013) Hispanic data point and the Asian/other datapoint were both matched to the White data point from the same study, and the Engelhardt et al. (2019) Hispanic data point, the Asian/other data point, and the multi-racial data point were all matched to the same White data point from their respective studies. An alternative approach would have been to enter the White data points two or three times, respectively, but this would have led to completely dependent data points in the database, which was not an acceptable alternative. So, our comparison between data points with ancestrally precise and imprecise group designations was not optimal but was maximally acceptable.

3.1.3. Discussion of results

As noted previously, there were no guidelines or generally accepted standards in the literature on the Scarr-Rowe effect for what constitutes a substantial difference in outcomes – Scarr did not supply them and neither did G&T. To answer the question of which differences between subgroups were substantial, we used Cohen's (1988) advice on the strength of effect sizes, namely $r = 0.10$ is small, $r = 0.30$ is moderate or modest, and $r = 0.50$ is large. As $r = 0.10$ is equivalent to $r^2 = 0.01$, $r = 0.30$ is equivalent to $r^2 = 0.09$, and $r = 0.50$ is equivalent to $r^2 = 0.25$, we made the choice of calling a $\Delta r^2 = 0.09$ moderate or modest, $\Delta r^2 > 0.09$ substantial, and $\Delta r^2 = 0.25$ large. Note, in Pesta et al. (2020) we used different rules of thumb.

For the moderator publication status, there were two moderate effects and two strong effects. However, the percentage of variance explained did not increase substantially, except for the case where two data points had highly similar outcomes. Except for E for Blacks, the amount of variance never increased substantially in the largest category (published studies). So, there was not a clear systematic moderator effect for publication status. Moreover, the substantially higher heritabilities for the two unpublished studies for both Whites and Blacks could be easily explained, as we used *g* scores that were based on a longitudinal study with data averaged across many waves. As Pesta et al. stated, this had the effect of reducing the nonshared environmental variance. Comparing White vs. Black heritability differences for the published and the non-published studies, the relative Δh^2 was 0.15. The Δh^2 was higher for the published studies, meaning that there was more evidence consistent with a Scarr-Rowe interaction in the published studies. This was a substantial, but not large, effect, that was in the direction that G&T seemed to predict.

Table 2
Results of outlier analyses for the black/interracial data point from the Minnesota Transracial Adoption Study by Scarr et al. (1993).

Variance component	Mean *rho*	SD_r	Distance from mean *rho*
A	0.60	0.151	−0.52
C	0.15	0.062	+0.81
E	0.25	0.116	+0.26

Note. The Black/interracial data point from the MTRAS was compared to all Black data points, including the MTRAS one.

For the moderator ACE estimates, there were two effects, but the explained variance did not increase substantially, except for negative ACE estimates for E for Blacks. So, there was not a clear systematic moderator effect for ACE outcomes. Comparing White vs. Black heritability differences, the relative Δh^2 was −0.01. The Δh^2 was lower for the positive ACE studies, meaning there was less of a Scarr-Rowe interaction in the samples with positive ACE estimates. This was not a large effect, but it was in the opposite direction to that which G&T seemed to predict.

For the moderator averaging within samples, there was one moderate effect and two strong effects. However, there were no substantial increases in the percentages of variance explained for most of the categories, except for the E component averaged within samples. Moreover, there was never increased variance explained for both categories. So, although there were some sporadic indications of a moderator effect, there was not a clear systematic moderator effect for the moderator 'averaging within samples.' Comparing White versus Black heritability differences, the relative Δh^2 was −0.16. The Δh^2 was lower for the 'not averaging within samples' studies, meaning that there was less of a Scarr-Rowe interaction in the samples where we did not average. This was a substantial, but not large, effect, but it was in the direction opposite to that which G&T seemed to predict.

For the moderator 'precision of group designation', there was just one modest effect, and there were no substantial increases in the variance explained. So, there was no moderator effect for ancestral precision of group designation. Comparing White vs. Black heritability differences, using ancestrally precise groups had no effect on the Δh^2. We could not make a comparison in the case of Whites vs. Hispanics plus others since we lacked an equivalent comparison. That said, the heritability for the combined Hispanic plus others group was higher than that for Whites, despite that group scoring lower on average (with most of the samples being Hispanic). Thus, our interpretation would not have changed had we categorized groups that way.

When there was a substantial effect for some of the six comparisons, it was not found in the other comparisons. So, there were no clear, systematic moderators that showed substantial effects for both Blacks and Whites for A, C, and E in all three cases, or at least not in the majority of the six comparisons for every potential moderator variable. It should also be taken into consideration that the outcomes for A, C, and E for one category were dependent: they must add up, within rounding error, to 1.00. Because of this dependence, one would expect, for instance, that a substantial difference between two categories for A would coincide with a substantial difference between categories for C or E. However, when there was a substantial effect for A, there were rarely substantial effects for C or E. Likely, some of the substantial differences between the squared correlations in the two categories were due in large part to chance. We conclude there is no clear indication of moderator effects for the four moderators derived from G&T's critiques.

Additionally, a finding which was potentially problematic for the hypotheses based on moderators derived from G&T was that the differences between the outcomes in the two moderator categories were not in the direction G&T would seem to have predicted for three of the four moderators. So, contrary to G&T's arguments, our methodological choices were not biased against finding Scarr-Rowe effects.

There were three limitations to these moderator analyses. First, the *Cochrane Handbook* (Higgins et al., 2019) stated that there should be at least ten data points for a meta-analysis and at least ten data points for a moderator analysis (see section 10.11.5.1). However, when the data points are not evenly distributed over the categories, even $K = 10$ is insufficient (see section 10.11.5.1). The data points for the moderators,

Table 3
Meta-analytical ACE outcomes for moderators.

	A				C				E			
SIRE	*K*	*rho*^{2}	Δ*rho*^{2}	% var.	*K*	*rho*^{2}	Δ*rho*^{2}	% var.	*K*	*rho*^{2}	Δ*rho*^{2}	% var.
Whites vs Blacks	30	0.59	0.02	2.13	28	0.18	0.05	14.12	28	0.25	0.01	10.08
Whites	15	0.58		2.25	14	0.20		10.62	14	0.24		17.21
Published	13	0.56	0.11	2.47	12	0.20	0.02	9.48	12	0.26	0.10	37.5
Non-published	2	0.67		78.6	2	0.18		159.9	2	0.16		242.8
Positive ACE estimates	12	0.58	0.12	1.92	11	0.19	0.07	10.17	11	0.24	0.04	15.42
Negative ACE estimates	3	0.46		18.22	3	0.26		14.11	3	0.28		33.30
Not averaged within sample	6	0.64	0.09	1.07	5	0.19	0.00	7.65	5	0.19	0.07	11.43
Averaged within sample	9	0.55		8.51	9	0.19		13.54	9	0.26		66.98
Blacks	15	0.60		2.0	14	0.15		26.69	14	0.25		7.14
Published	13	0.52	0.26	7.73	12	0.17	0.06	29.61	12	0.32	0.21	72.36
Non-published	2	0.78		1.54	2	0.11		142.5	2	0.11		6.20
Positive ACE estimates	11	0.60	0.11	1.70	10	0.15	0.06	28.37	10	0.25	0.07	5.39
Negative ACE estimates	4	0.49		5.32	4	0.21		26.93	4	0.32		35.66
Not averaged within sample	6	0.75	0.25	1.64	5	0.11	0.06	42.36	5	0.11	0.21	12.99
Averaged within sample	9	0.50		11.18	9	0.17		30.54	9	0.32		155.5
Whites vs Blacks	30	0.59		2.13	28	0.18		14.12	28	0.25		10.08
Precise group designation	15	0.58	0.02	2.25	14	0.20	0.05	10.62	14	0.24	0.01	17.21
Less precise group designation	15	0.60		2.00	14	0.15		26.69	14	0.25		7.14
Whites vs Hispanics + other imprecise^x	21	0.63		2.77	21	0.16		8.99	21	0.21		13.69
Precise group designation	9	0.60	0.11	4.12	9	0.18	0.01	8.79	9	0.23	0.06	13.28
Imprecise group designation	12	0.71		2.45	12	0.17		18.55	12	0.17		18.55

Note. K = number of data points; *rho*^{2} = mean meta-analytical value of, respectively, A, C, and E; Δ*rho*^{2} = *rho*^{2} of the first category minus *rho*^{2} of the second category; % var. = percentage of variance in the meta-analytical data points explained by sampling error. Other imprecise = Asians/others + multi-racial + non-White + non-White MTRAS. All the scores of Δ*rho*^{2} are reported as absolute numbers because we have no a priori theory on how to compare the outcomes.

publication status and ACE outcomes, were very unevenly distributed, so the moderator outcomes should be interpreted with care. It should be noted also that the data points for the other two moderators were evenly distributed.

Second, Pesta et al. (2020) ran three moderator analyses, and the present authors derived four moderators from G&T, so all in all, no less than seven moderators were being tested on the same meta-analytic dataset, so there is a strong risk of capitalization on chance (Hunter & Schmidt, 2004). We cannot therefore rule out the possibility that quite a few of the substantial differences between categories in either study were Type I errors.

Third, the Cochrane Handbook (section 10.11.5.4) stated that researchers should ensure a scientific rationale for investigating each characteristic for a moderator. However, G&T would have us instead 'throw everything against the wall and see what sticks.' They repeatedly stated that a specific approach taken by Pesta et al. was flawed, but they did not show any instances of studies where two methodological or statistical approaches were compared, yielding substantially different outcomes. G&T should have formulated clear hypotheses on the statistical and methodological flaws in Pesta et al. Another limitation of this study was that the moderators had not been sufficiently embedded into the literature. These three clear limitations made it even less plausible that the variables suggested by G&T were true moderators in this meta-analysis.

3.2. Power analysis

G&T noted we did not run power analyses. We followed Scarr's own advice in this regard (Scarr-Salapatek, 1973, *pp*. 1045–6). However, we agree that power analyses could have added to the study and so we have included them here in Appendix D. They count against the idea that our results were concerning. Regardless, G&T should not have claimed that our analyses were "underpowered." It is not possible to claim something is underpowered without qualification because there is no absolute sense in which a study can be underpowered. A study is only ever underpowered to detect an effect of a specific magnitude. And, as noted above, proponents of Scarr-Rowe interactions have not yet quantified their expectations.

4. G&T's call for censorship and attack on *Intelligence*

4.1. G&T's vision of science as compared with Sandra Scarr's

G&T thought they were providing "a cautionary tale" (*p*. 12) about "racist and eugenicist behavioral-genetics research" (*p*. 12), carried out by "individuals with less than honorable intent" (*p*. 12). They also claimed that Pesta et al. have published "problematic work in traditional subscription journals" (*p*. 12). Regarding recommendations going forward, G&T called for censorship. They proposed that "mainstream academic journals that wish to be taken seriously should not be bulletin boards on which anyone can post a study regardless of its rigor, ethics, or underlying motivation" (*p*. 12), and that editors need to "protect the integrity of the journal" (*p*. 12).

Moreover, G&T asserted that screening articles is "especially important for inflammatory and potentially harmful topics such as *alleged* group differences in intelligence" (*p*. 12, emphasis added). G&T next asked: "What can be done to discount low-quality, potentially harmful research…" (*p*. 13). Thereafter, they answered their own question: "a more active response is needed, involving interdisciplinary coordination at multiple levels from publishers to editors, editorial boards, peer reviewers, and promotion committees" (*p*. 13), and that "Publication outlets must develop protocols for responding to problematic content housed on their platforms" (*p*. 13). Comments like these have been addressed elsewhere (Haier, 2020).

G&T's approach to studying race and ethnic group differences in intelligence radically differs from Sandra Scarr's. Regarding her genetic-admixture study, Scarr (2009) noted: "My colleagues and I discussed the potential implications of the study and were prepared to report the results, whatever they were…Excluding minorities from mainstream research for fear of unpopular findings is not an option for reputable science… (para, 18).” Further on the issue of censorship, Scarr (1981, *pp.* 531–2) stated:

> On the third point—the possible danger to society of knowledge about genetic differences in behavior—my position is unequivocal. In my view, there is no danger so great as the suppression of knowledge. There is nothing we could learn about ourselves that would justify abridgment of scientific inquiry. There are methods of investigation that pose unconscionable threats to the participants in research. Methods should be subject to regulation. But there should be no regulation of scientists' rights to think, propose, and conduct ethical investigations on any question, however distasteful it might be to others…We should all tremble if the true believers of one position were to gain the power to silence dissent. And so, I do not believe that ideas are dangerous, however misguided and outlandish they may seem to me, but I quake at the self-appointed guardians of any orthodoxy.

4.2. G&T's attack on Intelligence

In their attack on our investigation of Scarr and Rowe's Original Hypothesis**,** G&T also called out the journal *Intelligence,* its editors, and its reviewers, for not rejecting our work:

> We believe that Pesta et al. (2020)'s publication in a mainstream journal represents a failure of the editorial and peer review process. We find it hard to imagine that any qualified, non-partisan intelligence researcher or behavior geneticist who reviewed the paper in sufficient depth would deem it worthy of publication.
>
> We have chosen to distance ourselves from *Intelligence* by no longer submitting or reviewing manuscripts there until substantial changes to the journal's editorial policies and practices have been made.

To preview content from Appendix E, we reacted here by carefully reading G&T's critique, while meticulously evaluating all their criticisms. Above, we documented frequent, false, inaccurate, and potentially misleading claims in G&T's article. Next, we submitted an early version of our manuscript to *Perspectives on Psychological Sciences* (PoPS), which is the journal where Giangrande and Turkheimer (2022) published their article about our work (note that G&T neither contacted us nor the Editor of *Intelligence*, with the intention to submit a critique at *Intelligence*; see, however, the policies of the journal, *Intelligence,* Haier, 2020)).

The editor of PoPS (Klaus Fiedler, at the time) solicited reviews from fully 19 researchers. According to Fiedler, an unprecedented 17 of them declined to review our rebuttal. Of the two researchers who ultimately reviewed our manuscript, one was very positive, whereas the other was very negative. Based on these reviews, Fiedler rejected our manuscript, but left the door open for us to submit another new article as our rebuttal. We took him up on his offer.

Subsequently, Fielder stated that he sent the second version of our rebuttal to prominent researchers, who all recommended rejection. Sometime later, though, Fiedler admitted to APS (as part of our appeal, see below) that he never sent the manuscript out for review because he deemed that it was not sufficiently interesting to PoPS' broad reader base. We believe, however, that Fiedler unambiguously violated the ethical guidelines of the Committee on Publication Ethics' (COPE; of which PoPS is a member). COPE guidelines clearly state that “**Authors of criticised material should be given the opportunity to respond** (emphasis added),” and that “**Errors, inaccurate, or misleading statements must be corrected promptly and with due prominence** (emphasis added).”

As part of our appeal, we also invited professor Turkheimer to contribute a rebuttal piece, such that our articles would appear back-to-back in the same issue of PoPS. However, we heard back from neither Turkheimer nor Fiedler on this point. We therefore decided to file formal ethics complaints with both APS and Sage Publishing, but APS deemed that we did not have a right to reply in their journal, and Sage has yet to respond. Finally, we decided to submit our rebuttal here at *Intelligence*, and we have detailed all the steps we took with PoPS in Appendix E.

5. Conclusion

Here we documented G&T's various erroneous claims about the literature on the effects Scarr and Rowe proposed; the misrepresentations of our methods and conclusions; the isolated demands for rigor and inconsistencies, and their uninformed statements about meta-analysis. Moreover, G&T never empirically tested any of these claims, although the relevant data were readily available to them, as contained in our meta-analysis.

For example, and despite cautions contained in our meta-analysis, G&T seemingly obsessed over how our Heritability × Group difference results “unequivocally” contradicted our conclusion (but see our disclaimer above for why we appropriately concluded that our results were not unequivocal on this issue). Another example includes G&T's proposed treatment of outliers and moderators in the meta-analysis. We found in general almost no empirical support for any of G&T's claims here, especially the insinuation that we were systematically biased against finding Scarr and Rowe's effect. Again, however, G&T could have easily backed their speculations up in their critique, simply by analyzing already available data.

G&T also seemed to have framed their critique on a misstatement of Scarr's research. For example, it was clearly the case (as revealed by our literature review above) that Scarr was interested in Race/ethnicity x Heritability interactions, independent of SES, and it was clearly the case that recent studies on Gene × Environment interactions have also examined Race/ethnicity x Heritability interactions. Moreover, in a paper where the term “Scarr-Rowe interaction” was coined, Turkheimer himself (Turkheimer et al., 2009) argued that researchers should follow Scarr-Salapatek (1971a), by re-analyzing all heritability studies with variance components moderated by socioeconomic status or age or gender or race. Frankly, G&T's claim that the Scarr-Rowe interaction has nothing to do with race (independent of SES) is simply false.

Additionally, G&T expressed concern about the possibility of error in our race/ethnicity categorizations, when in fact they made theoretical sense given the hypotheses under investigation. Our categorizations were consistent with how Scarr, Rowe, Guo, and others have grouped individuals. As an example, G&T criticized our inclusion of ancestrally heterogeneous groups, but Van Den Oord and Rowe (1997) and Guo and Wang (2002) did likewise for similarly sound reasons.

Next, G&T seemingly misconstrued many of our statements and conclusions. For example, they claimed we: (1) argued that the results confirmed a genetic hypothesis when we did not; (2) did not provide a sound rationale for our expectation that higher-scoring Asians should have higher heritabilities than Whites, but we did; (3) dismissed additional negative ACE estimates, when the relevant samples did not have negative ACE estimates; (4) inappropriately used noncognitive scores from Mollon et al. (2021) when in fact we clearly stated that we used g scores. Note that many of G&T's other criticisms also apply to Tucker-Drob and Bates' (2016) meta-analysis of SES x Heritability interactions (which also noted potential confounding by race/ethnicity), yet this study went uncritiqued in that regard.

Finally, we understand that testing aspects of heritability differences is complex and controversial. We are open to criticism, but critics have an obligation to present their views professionally and within the norms of scientific debate. We strived to do this here.

Data availability

Data will be made available on request.

Appendix A. List of heritability comparisons for variables besides intelligence

Table A1 presents a list of heritability comparisons by Race/ethnicity for variables besides intelligence.

Table A1
Racial/ethnic heritability comparisons for variables besides intelligence.

Author	Group designations	Specific Groups	Variables
Biagini et al. (2022)	Race	White and Black American	Asthma risk, exposures to secondhand smoke, and traffic-related air pollution
Khan et al. (2019)	Race	White and African American	Cardiac mechanics (e.g., global circumferential strain)
Kolifarhood et al. (2019)	Ethnicity	Europeans, Mexicans, Middle Easterners, Asians, Africans, Latinos, Hispanics, and American Indians	Blood pressure traits
Enkhmaa, Anuurad, Zhang, Kim, and Berglund (2019)	Ethnicity	Caucasian and African American	Apolipoprotein-A traits
Polubriaginof et al. (2018)	Race and ethnicity	White, Hispanic/ Latino, and Black/AA	Height
Musani et al. (2017)	Racial/ ethnic groups	White and Black	Metabolic syndrome
Gusev et al. (2016)	Ancestries	European and African Americans	Prostate cancer
Bares, Kendler, and Maes (2016)	Race	White and African American	Cigarette smoking
Munn-Chernoff et al. (2015)	Race/ethnicity	European and African Americans	Major depressive disorder and overeating/binge eating
Sartor et al. (2013)	Ethnicity	European and African Americans	First drink and problem alcohol use

Authors' terminology was used.

Appendix B. ACE estimates with and without negative values

Table B1 presents ACE estimates from Pesta et al. (2020) for those samples with any negative ACE values. Estimates are presented with negative values set to zero ("Without Negatives") and, alternatively, with the negative values not set to zero ("With negatives"). Meta-analytic means were computed for the two sets of values. These were generated using SPSS 28 with a random-effects model, the Hunter-Schmidt estimator, and the inverse standard error of heritability as weights.

Table B1
ACE estimates from Pesta et al. (2020).

			Without Negatives			With negatives			
Sample	Race / ethnicity	*N*h	S.E. A	A	C	E	A	C	E
Philadelphia School Sample	Black	448	0.29	0.31	0.32	0.37	0.27	0.34	0.38
The Twin Study	Black	116	0.29	0.59	0.13	0.28	0.66	0.04	0.30
Philadelphia Twin Study	Black	154	0.24	0.48	0.11	0.41	NA	NA	NA
MIDUS	Black	10	0.85	0.94	0.00	0.06	1.84	−0.96	0.12
Florida Twin Study	Black	167	0.16	0.89	0.00	0.11	1.04	−0.17	0.13
Hunter-Schmidt Meta-analytic mean				0.68	0.09	0.24	0.71	−0.02	0.25
Philadelphia School Sample	White	230	0.33	0.28	0.44	0.28	0.22	0.48	0.30
The Twin Study	White	299	0.16	0.54	0.28	0.18	0.56	0.25	0.19
Philadelphia Twin Study	White	208	0.23	0.53	0.03	0.43	0.58	0.01	0.41
MIDUS	White	586	0.11	0.14	0.46	0.4	NA	NA	NA
Florida Twin Study	White	465	0.10	0.80	0.02	0.18	NA	NA	NA
Hunter-Schmidt Meta-analytic mean				0.48	0.23	0.28	0.48	0.23	0.28

Note. Samples without negative estimates are marked as NA; when creating average estimates, we used the estimates without negatives in place of NA. Summary means were generated using a random effects model, the Hunter-Schmidt estimator, and the inverse standard error of heritability as weights. Data were analyzed with SPSS 28.

Appendix C. Updated results from heritability × group difference analyses

Table C1 presents updated Results from Heritability × Group Difference Analyses. In Pesta et al. we reported heritability × group difference results using, as noted, the Satterthwaite approximation of the pooled error for heritability as weights. In retrospect, this was a conceptual error, since this method placed more weight on less reliable results. Moreover, the analysis for All comparisons involved dependency among comparison – e.g., the effects for the Black-Hispanic comparison are non-independent from the effects for the Black-White and White-Hispanic comparisons. Thus, we redo the analyses and include a category for (more or less) independent comparisons before discussing the results. For the All independent comparisons analyses, we use only the 26 White-non-White comparisons, since these only reuse the White heritability estimates in computing difference scores.

These updated results, based on least squares regression (using SPSS 28), are shown in Table C1. In this table, Model 1 shows the results for the bivariate association between Δh^2 and *d*. Model 2 adds only three covariates: age, test type, and estimation methods, along with dummy variables for the group comparisons (e.g., Black-White = 1, else = 0) in case of the analyses across groups. The former were the variables used as moderators in the Pesta et al. meta-analysis. Moreover, they were examples of specific potentially confounding variables suggested by Pesta et al.: "Across samples, however, several confounds existed, such as differences in age, methods of estimated h^2, and differences in cognitive measures" (*p*.10). For robustness sake, we use three different weights: square root of the inverse S.E. A, square root of the harmonic *N*, and equal weights. We report *p*-values rounded to the hundredth place.

Table C1
Results from heritability × group difference analyses.

		Weight	Model 1		Model 2	
	N_s		β	*P*-value	B	*P*-value
Black-White	15	SQRT(1/SE)	0.76	*0.00*	0.60	*0.01*
		SQRT(*N*h)	0.74	*0.00*	0.58	*0.01*
		Equal	0.90	*0.00*	0.66	*0.02*
White-Hispanic	7	SQRT(1/SE)	0.66	*0.11*	0.15	*0.50*
		SQRT(*N*h)	0.67	*0.10*	0.20	*0.42*
		Equal	0.70	*0.08*	0.15	*0.46*
Black-Hispanic	7	SQRT(1/SE)	0.70	*0.08*	0.34	*0.12*
		SQRT(*N*h)	0.69	*0.09*	0.31	*0.17*
		Equal	0.65	*0.12*	0.50	*0.06*
Any Other	11	SQRT(1/SE)	0.21	*0.54*	0.52	*0.39*
		SQRT(*N*h)	0.20	*0.56*	0.49	*0.42*
		Equal	0.25	*0.46*	0.54	*0.36*
All comparisons	40	SQRT(1/SE)	0.32	*0.04*	0.36	*0.00*
		SQRT(*N*h)	0.33	*0.04*	0.35	*0.00*
		Equal	0.38	*0.01*	0.37	*0.02*
All Independent comparisons	26	SQRT(1/SE)	0.58	*,00*	0.36	*0.00*
		SQRT(*N*h)	0.60	*0.00*	0.36	*0.00*
		Equal	0.63	*0.00*	0.40	*0.00*

Note: Model 1 reports the β for analyses with d as the dependent and Δh^2 as the independent. Model 2 added the three moderator variables used in the meta-analysis: age (sample mean or median), test type (IQ or g vs. any other), and analysis type (Falconer's vs. any other). Model 2 also adds group comparison dummies for the All comparisons and the all-independent comparisons. *p*-values are rounded to the hundredth place.

As seen in both models, there is only a statistically significant effect (i.e., $p < .05$) for the Black-White and for the aggregate comparisons. Moreover, the three moderators have a noticeable effect on the White-Hispanic, Black-Hispanic, Any Other, and all independent comparisons coefficients, implying that these coefficients are not stable across factors that vary between samples as we had cautioned.

Nonetheless, judging from the analysis using all independent comparisons, the effects remain significant in the face of obvious confounders. Yet, given that these results control for only a few factors that differ between samples (in Model 2) and that the number of independent pairwise comparisons is small we would not describe these results as providing "unequivocal" evidence in support of a modified version of Scarr and Rowe's hypothesis as G&T do. Instead, we reiterate our conclusion: "When we looked across samples, we found evidence consistent with the interaction… however, several confounds [exist]" (*p.* 10).

Overall, we argue that examining heritability × group difference is an interesting possible alternative way to test for Environment x Heritability interactions. This method could be explored further in the future. But in no way do these results contradict our point about an absence of Race/ethnicity x Heritability interactions.

Appendix D. Power analyses

This section includes power analyses but requires theoretical explanation prior to their presentation. Firstly, there were two desired quantities a power analysis could provide us. One quantity is how well-powered we were to test the Scarr-Rowe effect and the other was how well-powered we were to detect any differences in heritability between groups. The second quantity corresponds to the minimum detectable group difference in heritability. The first requires theory and, we argue, explicit expansion of the Scarr-Rowe hypothesis into 'weak' and 'strong' forms.

The 'weak' form of the Scarr-Rowe hypothesis can be dubbed the 'SES-only' variant. Under the weak form of the Scarr-Rowe hypothesis, the expected difference in heritability between groups is the product of the Scarr-Rowe effect in a group and their SDs of difference in SES compared to another group.

The 'strong' form of the Scarr-Rowe Hypothesis is how Scarr-Salapatek, 1971a, *p.* 1294) originally formulated it: for Scarr, the effect of "disadvantage" was never localized to socioeconomic status. She did not believe that, which is why she was explicit that race can be associated with non-SES developmental insults. Conducting a power analysis for the "strong" or Scarr and Rowe's Original Hypothesis is not possible to do precisely. It is based on combining the effects of how much lower a group's SES is *and* how much worse its conditions are aside from SES. That would mean, for example, assessing the effect of racial discrimination or intergenerational poverty and distress on heritability. In virtually all scenarios, the strong Scarr-Rowe hypothesis predicts larger heritability differences than the weak one because the biometric variance components for lower-scoring groups ought to be lower for reasons besides SES alone.

The results of our power analyses are contained in Table D1. All analyses were conducted as comparisons between the target group (e.g., Black, Hispanic, etc.) and Whites, with the per-group *k* given as whichever group had a lower number of studies and the sample size used being the median harmonic *N* for the smaller group. The estimates are for the power to detect meta-analytic subgroup differences with the I^2 used being the upper-bound for a meta-analytic value for the meta-analysis with each pair of groups. All of these just mentioned methodological choices lead to reduced power; if we used the actual values, power would be considerably higher in each case. We only presented negative differences here because the expectation based on Scarr's foundational work was that lower scoring non-White groups would have lower heritability values, but power for this analysis was higher in the positive direction. The minimum detectable effect (MDE) was conceptualized not as the lowest-possible significant effect, but the lowest effect that would provide 80% power at $\alpha = 0.05$.

The expected size of the weak Scarr-Rowe effect was not calculated for the "Other" group because that grouping was a mix of higher and lower scoring groups. The effect size for the Black group was based on SES values provided by Warne (2021), while the SES difference was estimated from 0.4

to 1 *d* for the Hispanic group to simulate a variety of values. Both utilized the variously moderated values for the Scarr-Rowe effect from Tucker-Drob and Bates (2016).

Table D1
Power to detect heritability differences in Pesta et al. (2020).

Group	*−5%*	*−10%*	*−20%*	*−30%*	*−40%*	Post Hoc	*MDE*	*Weak Scarr-Rowe**	*Observed Δ*
Black (k = 15)	60%	99%	99%	99%	99%	21%	−6.5%	42–94%	3%
Hispanic (k = 7)	18%	54%	98%	99%	99%	69%	−13.7%	8–72%	11%
Other (k = 4)	7%	12%	32%	60%	86%	24%	−37.6%	Not Computed	−17%

Note. * The Scarr-Rowe effect size is given as a range.

As can be seen in Table D1, the a priori power to detect the Scarr-Rowe effect, given the conservative methodological choices noted above, ranged from meager to considerable in the comparison with the Black group, and from nearly-nothing to nearly-sufficient in the Hispanic group. However, both groups' results ran in the opposite direction of the predictions from the weak Scarr-Rowe effect. Despite being socioeconomically disadvantaged and — perhaps — the targets of race-based discrimination and prejudice, both groups had *higher* heritability than the White group. Using our smallest value for the Scarr-Rowe effect (*$a' = 0.060$)*, the Black group was expected to have 3.9% lower heritability, but instead, they had 2.5% higher heritability, for a cumulative difference between the predictions of the weak Scarr-Rowe hypothesis and the meta-analytic difference of 6.4%. There was 83% power to detect that difference. Using the lowest level of SES differentiation (0.4 *d*) and the same small Scarr-Rowe effect magnitude for the Hispanic group, the expected level of heritability depression was 2.4%, but the observed effect was 10.6% higher heritability, and thus there was 84% power to detect a difference between observations and weak Scarr-Rowe effect predictions. If we instead used upper-bound levels of SES differentiation for the Hispanic sample and used the upper-bound Scarr-Rowe effect ($a' = 0.123$) for both, we were powered at 99% for both differences. Depending on how one qualifies the size of the Scarr-Rowe effect and racial/ethnic socioeconomic disparities, power is substantially affected.

However, these empirical values for power to detect the weak Scarr-Rowe effect were interpretable the same way as post hoc power is more generally, which is to say that they were recapitulations of the *p*-values for our observed differences. The data *did* deviate strongly from the predictions of the SES-only version of the Scarr-Rowe hypothesis.

To provide a further test of how much sampling error might have affected our results, the original meta-analysis was redone with all standard errors set to the maximum standard error in our dataset, 0.92, for convenience. It would not make a difference if they were all set to 0.01, since they would all be weighted the same as a result. Observed differences in heritability declined across the board, to 2.9%, 6.9%, and 13.7% relative to the White group. Our post hoc power for tests of the Scarr-Rowe hypothesis increased for the Black group, to 85%, while dropping in the Hispanic group, to 52%. If we used the largest rather than the smallest Scarr-Rowe effects and levels of Hispanic SES differences, the power was 99% for the Black group and 98% for the Hispanic one.

Next, we tested moderation by 'problematicness'. We dummy coded a variable indicating whether G&T wrote that a given study or sample was suspect and assessed whether this moderated our results. We also coded a version of this variable multiplicatively such that if a study was criticized twice, it received a value of 2, and if it was criticized once, a value of 1, etc. Whether coded categorically or continuously, problematicness had no significant effects and did not affect other results, and, moreover, a model with it did not fit better than a model without it, as assessed by ANOVA.

Appendix E. Summary and timeline of events regarding our initial submission to *Perspectives on Psychological Science* (PoPS)

1. **Brief Summary**:

   Below we provide dates and email chains regarding our experience with PoPS. In brief, we submitted our original commentary to the journal, which the Editor (Klaus Fiedler) then sent out for review. The editor rejected our commentary but did leave the door open for us to submit a substantially revised version of our paper as a resubmission. The editor, however, desk-rejected our resubmission.

   We then appealed the editor's decision, by following the chain of command at both APS and Sage Publishing. While at least some of the entities we appealed to were responsive and informative, the bottom line is that PoPS will not afford us the opportunity to publish any reply to G&T's scathing critique of our meta-analysis (published in *Intelligence* in 2020).

2. **Timeline of Events**:

1. APS publishes a critical review of us, written by Giangrande and Turkheimer (2022). Then-editor Dr. Laura King neither asked any of us to peer-review G&T's article (as is convention), nor did she invite any of us to submit a reply (as is convention).
2. We emailed Dr. King asking whether we could submit a reply to G&T's article. Instead of acting immediately and giving us a chance to quickly react, Dr. King took approximately half a year to respond to our email, at which time Dr. Fiedler had replaced her as editor. Dr. King informed us that we should indeed contact Dr. Fiedler regarding our request.
3. On **3/8/2022**, we submitted our first "Commentary" regarding Giangrande and Turkheimer (2022) to PoPS. Editor Klaus Fiedler emailed us his decision on **4/15/2022.**
4. Based on Professor Fiedler's feedback, we substantially revised our manuscript and sent it back to PoPS on **6/1/2022**. We received Fiedler's rejection email on **6/3/2022**. Professor Fiedler stated that his decision was based on review replies from renowned expert reviewers. The reviewer comments were not attached to Fiedler's email. Professor Fiedler subsequently told APS that he did not send the revised manuscript out for review. The resubmission and decision letters appear in Section 3.1 below.
5. We contacted Professor Turkheimer on **6/21/22** to see if he would be willing to write back-to-back replies for PoPS. However, we received no reply from Professor Turkheimer. This letter is shown in Section 3.2.
6. In **August of 2022** (exact date unknown), we contacted Douglas Detterman, the founding editor of the journal *Intelligence*, to see if he would write a letter of support for us, addressed to Professor Fiedler. Professor Detterman subsequently sent a letter urging Professor Fiedler to allow us a reply to Giangrande and Turkheimer (2022). Professor Detterman's letter is shown in Section 3.3.

7. We submitted an appeal to the Association for Psychological Science (APS) on **9/22/2022**, receipt of which was acknowledged on **9/23/2022**. APS replied on **10/5/2022**, deeming that PoPS did not have an obligation to publish a reply because our original article was published in *Intelligence*. On **10/14/2022** we submitted a reply letter asking for clarification since (1) the caveat about where the original paper was published is not noted anywhere in COPE's ethics guidelines, and (2) our other two complaints were ignored. However, APS never responded. A part of the correspondence with APS is shown in Section 3.4.
8. We submitted an appeal to Sage Publishing on **10/19/2022**, receipt of which was acknowledged on **10/20/2022.** In that notice of receipt, Sage Publishing noted that "we cannot give you periodic updates on this situation, but we will inform you of any outcomes after a decision has been made." We did not receive a reply afterwards. The appeal letter to Sage Publishing is virtually identical to the appeal letter sent to the APS.

3. **Supporting documents**

3.1 Our Resubmission letter to PoPS and subsequent rejection letter.

3.1.1 Our resubmission letter to PoPS

June 1, 2022

Klaus Fiedler.

Editor, *Perspectives on Psychological Science.*

Dear Professor Fiedler:

You had previously rejected a version of this manuscript in April of 2022. However, you mentioned that "the door is open for your submitting a new manuscript that deals with the G&T critique in a convincing scientific style, based on original evidence or on an upfront discussion of methodological problems that are of interest for a broader readership, not just for a few personally involved authors." We hope we have done this here and ask that the new version please be reviewed for possible publication in PoPS.

Additionally, we felt we were obligated to address the "reviewer concerns" stemming from our original submission (especially given your difficulty finding reviewers here, which makes it possible the old reviewers might get this version as well). So, appended below, please find our response to the reviewers from our original manuscript.

Sincerely,

Bryan Pesta and Co-authors

3.1.2 Decision letter from PoPS

From: Perspectives on Psychological Science <Email Address>

To: <Bryan Pesta>

Sent: Fri, Jun 3, 2022 4:51 am

Subject: Perspectives on Psychological Science - Decision on Manuscript ID PPS-22-163

03-Jun-2022

Dear Dr. Pesta:

Thank you for submitting your manuscript # PPS-22-163 entitled "On Group Differences in the Heritability of Intelligence: A Reply to Giangrande and Turkheimer (2022)" to Perspectives on Psychological Science. I was happy to be offered the help of some of the most renowned expert reviewers for the evaluation of this manuscript. Please find their comments below.

Based on the reviewers' advice and on my own careful reading of your manuscript, I have concluded that this manuscript is not suitable for publication in Perspectives. I am therefore sorry to write that I have decided to decline the manuscript.

I realize that authors such as yourself work hard on the preparation of these manuscripts, and I am sorry that I cannot bring you better news. We hope that the reviewers' constructive feedback will assist you in seeking publication elsewhere.

Thank you for considering Perspectives on Psychological Science for the publication of your research. I hope the outcome of this specific submission will not discourage you from the submission of future manuscripts.

Sincerely,

Klaus Fiedler

Editor, Perspectives on Psychological Science

Reviewer(s)' Comments to Author: [note, the email ends here]

3.2 Request to participate in back-to-back reply in PoPS sent to Professor Turkheimer

June 21, 2022

Dear Professor Turkheimer,

You are likely aware of a pre-print colleagues and I published on PsyArxiv (https://psyarxiv.com/qbkcg/). This represents our reply to your 2022 article with Evan Giangrande (in *Perspectives on Psychological Science;* PoPS).

The causes of racial and ethnic group differences in intelligence are important scientific topics, and we believe they require full and thorough scientific discussion. We therefore think it would interest PoPS readers to see our reply, back-to-back with your (i.e., Giangrande & Turkheimer's) rebuttal of it.

In line with this, the *Best Practice Guidelines on Publishing Ethics* of Wiley states that:

"If an item of correspondence discusses a specific article, the journal should invite the authors of the work to respond before the correspondence is published. When possible, the correspondence and the authors' response should be published at the same time."

To advance the discussion, we've taken the initiative here to see if you would be interested in contributing the rebuttal part to a back-to-back reply in PoPS. If you are willing to do this, would you please indicate so by replying here at your earliest convenience?

Sincerely,

Bryan Pesta and Coauthors

3.3. Prof. Detterman's letter to Prof. Fiedler

Prof. Fiedler,

I am writing you as the new editor of *Perspectives on Psychological Science*. Much of what I will be discussing happened before you became editor and I understand you had nothing to do with it. The article I will be addressing is: Giangrande, E. J. and Turkheimer, E, (2022). Race, ethnicity, and the Scarr-Rowe Hypothesis: A cautionary example of fringe science entering the mainstream. *Perspectives on Psychological Science*, 17(3), 696–710.

Before I get to my main point, let me tell you a little about myself. I edited the journal *Intelligence* from its founding for nearly 40 years. I have had to deal with many controversial papers. I must say that the Giangrande and Turkheimer paper was one of the most malicious scientific papers I have ever read. But I will put that aside because I know you had nothing to do with that. What I am concerned with is the science it addresses.

I am writing because the main target of the Giangrande and Turkheimer paper was a paper by Pesta et al. (2020). Pesta et al. submitted a rebuttal to the Giangrande and Turkheimer article but it finally was rejected without review by your journal. It is grossly unfair to not allow authors who were attacked so viciously in a paper including both scientific and ad hominem declarations to respond to the scientific arguments made in the original paper. It also seems unfair to your readership to allow only one side of an argument to be presented. If you are concerned about giving Pesta and co-authors the last word you could give Giangrande and Turkheimer a chance to reply to them.

From the December 2021 interview with you on the APS website I read, you are seeking to present the full range of psychological science. It seems to me you must allow these authors to submit a rebuttal in the name of fairness and good science. In my opinion, to do less would be reprehensible.

Sincerely,
Doug
Douglas K. Detterman
Louis D. Beaumont University Professor Emeritus
Department of Psychological Sciences
Case Western Reserve University
Cleveland, OH 44106
216-395-4747 Office
216-287-7546 Mobile

3.4. Correspondence with APS

3.4.1 Initial email to APS

September 23, 2022

Thu, Sep 22, 2022 8:31 am
<Bryan Pesta +2 more>
September 22, 2022
Dr. Alison Gopnik, President
Association for Psychological Science

Re: Three violations of ethical rules by journal editors for recent editorial decisions at APS (PoPS, PPS-22-163).

Dear Dr. Gopnik,

My name is Bryan Pesta. I am a research psychologist located in Ohio. I am writing this email on behalf of myself and my colleagues (signed below). We write to file a formal complaint regarding three ethics violations committed by APS editors. Specifically, our complaint regards the behaviors of both the former (Dr. Laura King) and current (Dr. Klaus Fiedler) editors of the APS journal, *Perspectives on Psychological Science* (PoPS).

This incident is detailed below. I hope that you would please consider our appeal, as the PoPS editors here clearly violated both scientific convention and the Committee on Publication Ethics' (COPE) ethical guidelines. We note also that PoPS is indeed a member of COPE.

We argue that both the present and the previous editor of PoPS broke three rules regarding the ethical behavior of editors. Specific details of the incident are appended below.

Sincerely,
Bryan J. Pesta,
Emil Kirkegaard,
Jan te Nijenhuis,
Jordan Lasker,
John Fuerst

Note The articles relevant to our complaint include:

Pesta, B. J., Kirkegaard, E. O. W., te Nijenhuis, J., Lasker, J., & Fuerst, J. G. R. (2020). Racial and ethnic group differences in the heritability of intelligence: A systematic review and meta-analysis. Intelligence, 78, 101408

*Giangrande, E. J., & Turkheimer, E. (2022). Race, ethnicity, and the Scarr-Rowe Hypothesis: A cautionary example of fringe science entering the mainstream. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916211017498

*The vitriolic nature of this article quickly becomes clear when briefly skimming through it.

Our rebuttal of Giangrande & Turkheimer's review (rejected by Dr. Fiedler):

Pesta, B., te Nijenhuis, J., Lasker, J., Kirkegaard, E., & Fuerst, J. (2022). On group differences in the heritability of intelligence: A reply to Giangrande and Turkheimer (2022)

Timeline

(1) APS publishes a critical review of us, written byGiangrande and Turkheimer (2022)

Their review focused on an article we published in the journal, *Intelligence* (Pesta et al., 2020). G&T's article is patently inflammatory. For example, at one point in the article, G&T called for APS readers to boycott the journal *Intelligence*, just because they published our paper.

Notably, then-editor Dr. King neither asked any of us to peer-review G&T's article (as is convention), nor did she invite any of us to submit a reply (as is convention). These actions clearly violate ethical guidelines for journal editors.

(2) Our request to rebut G&T is sent to Dr. King (2021)

We emailed Dr. King asking whether we could submit a reply to G&T's article. Instead of acting immediately and giving us a chance to quickly react to the ubiquitous inaccuracies and misleading statements, Dr. King took approximately half a year to respond to our email, at which time Dr. Fiedler had replaced her as editor. Dr. King informed us that we should indeed contact Dr. Fiedler regarding our request. In sharp contrast, Dr. Fiedler reacted quickly and thereafter encouraged us to submit our rebuttal.

(3) Turkheimer attacks us on Twitter

Dr. Turkheimer attacked Pesta and co-authors on Twitter, calling them racist and anti-Semites.

(4) Our first submission is sent to PoPS (2022)

Dr. Fiedler approved our request to submit the reply. However, he ultimately rejected our submission after receiving comments from two reviewers. Note that one of these reviewers was highly positive regarding our submission, whereas the other was highly negative.

Further, highlighting the controversial nature of our research topic, Dr. Fiedler wrote:

My apologies first of all for the extremely long delay of my editorial feedback. I have to say that in almost 30 years of editorial work, I never experienced that as many as 17 invited reviewers declining …

Nonetheless, Dr. Fiedler left the door open for us to submit a new article (versus a revision):

To conclude, for the reasons summarized in this letter, I have to decline the possibility of publishing this manuscript in PPS. I hasten to add that the door is open for your submitting a new manuscript that deals with the G&T critique in a convincing scientific style, based on original evidence or on an upfront discussion of methodological problems that are of interest for a broader readership, not just for a few personally involved authors.

Note that Dr. Fiedler did not address his ethical obligation of allowing critiqued authors to defend themselves via rebuttal (as per COPE guidelines). Rather, he used "reader interest," and the lack of new data in our submission (which was a mere reply, versus a research article) as the reasons for his editorial decision.

Nonetheless, given Dr. Fiedler's feedback, we made substantial changes to the original manuscript, with an eye toward toning it down while also addressing all reviewer concerns.

(4) Our revised reply is submitted to PoPS

Dr. Fiedler asked us to submit the revision as a "new paper." His request, however, was misguided, as ours was clearly not a "new paper" but a revision of a previously submitted manuscript. We believe that by asking us to submit a new article, Dr. Fiedler could thereby bypass the reviewers for our original submission (including the very positive reviewer).

We nonetheless submitted our revision to PoPS on June 1st, 2022. The revised manuscript was rejected on June 3rd, 2022—just three days later. Cleary, Dr. Fiedler desk rejected our revised manuscript. However, in the decision letter, Dr. Fiedler mentioned that several "renowned expert reviewers" were consulted, and that their comments can be found at the bottom of his letter. The comments, though, were not included in the email. To wit:

Thank you for submitting your manuscript # PPS-22-163 entitled "On Group Differences in the Heritability of Intelligence: A Reply to Giangrande and Turkheimer (2022)" to Perspectives on Psychological Science. I was happy to be offered the help of some of the most renowned expert reviewers for the evaluation of this manuscript. Please find their comments below.

Based on the reviewers' advice and on my own careful reading of your manuscript, I have concluded that this manuscript is not suitable for publication in Perspectives. I am therefore sorry to write that I have decided to decline the manuscript.

We deem it unlikely that our long, detailed revision, together with our lengthy Supplementary Materials file, could be reviewed by reviewers in such an extremely short time. This becomes even more unlikely when we take into consideration that no less than seventeen reviewers declined to review the first version of the manuscript, leading to a review procedure of several months.

Thereafter, we made follow-up requests to both the PoPS editorial system (twice) and to Dr. Fiedler asking for the reviewer comments that were referenced above. We received no reply to any of these. Ignoring our repeated requests to supply us with the alleged reviewer comments strengthens our impression that Dr. Fiedler did not send the manuscript out for review.

(5) We file a formal appeal of Dr. Fiedler's editorial decision

In response to Dr. Fiedler's second rejection of our manuscript, we started an appeal wherein we followed the chain of command up through contacting the APS leadership team, which is where we are today. Our appeal included the following steps:

We first emailed Dr. Turkheimer, inviting him to participate in a traditional scientific debate, where Pesta and co-authors would reply to Giangrande and Turkheimer (2022), and where Giangrande and Turkheimer would then write a reaction. The plan was to publish the two papers in the same issue of PoPS. Dr. Turkheimer, however, did not reply to our request.

We next asked Douglas Detterman, the founding editor of the journal *Intelligence*, to write to Dr. Fiedler requesting that he allow us to publish our rebuttal. As of today, Dr. Fiedler has yet to reply to Dr. Detterman. Professor Detterman wrote:

Prof. Fiedler,

I am writing you as the new editor of *Perspectives on Psychological Science*. Much of what I will be discussing happened before you became editor and I understand had nothing to do with it. The article I will be addressing is: Giangrande, E. J. and Turkheimer, E, (2022). Race, ethnicity, and the Scarr-Rowe Hypothesis: A cautionary example of fringe science entering the mainstream. *Perspectives on Psychological Science*, 17(3), 696–710.

Before I get to my main point, let me tell you a little about myself. I edited the journal *Intelligence* from its founding for nearly 40 years. I have had to deal with many controversial papers. I must say that the Giangrande and Turkheimer paper was one of the most malicious scientific papers I have ever read. But I will put that aside because I know you had nothing to do with that. What I am concerned with is the science it addresses.

I am writing because the main target of the Giangrande and Turkheimer paper was a paper by Pesta et al. (2020). Pesta et al. submitted a rebuttal to the Giangrande and Turkheimer article but it finally was rejected without review by your journal. It is grossly unfair to not allow authors who were attacked so viciously in a paper including both scientific and ad hominem declarations to respond to the scientific arguments made in the original paper. It also seems unfair to your readership to allow only one side of an argument to be presented. If you are concerned about giving Pesta and co-authors the last word you could give Giangrande and Turkheimer a chance to reply to them.

From the December 2021 interview with you on the APS website I read, you are seeking to present the full range of psychological science. It seems to me you must allow these authors to submit a rebuttal in the name of fairness and good science. In my opinion, to do less would be reprehensible.

Sincerely,
Doug
Douglas K. Detterman
Louis D. Beaumont University Professor Emeritus
Department of Psychological Sciences
Case Western Reserve University
Cleveland, OH 44106
216–395-4747 Office
216–287-7546 Mobile

We wrote to Dr. Fiedler, asking him to allow us to react to G&T, referring to the COPE guidelines. We have yet to receive a reply from Dr. Fiedler. We therefore send our final appeal to the APS leadership team (this document).

Ethics Violations

Various guidelines exist regarding the ethical behavior of journal editors. Examples include COPE (2011); Wiley (2014), and the International Committee of Medical Journal Editors (ICMJE; 2022). We cite these sources below to derive three rules of the ethical behavior of editors that we believe were violated in our case.

Rule 1: Authors of criticized material should be actively invited to respond and should be given the opportunity to respond.

COPE (2011):

14.1 Editors should encourage and be willing to consider cogent criticisms of work published in their journal.

14.2 Authors of criticized material should be given the opportunity to respond.

Best practice for editors would include being open to research that challenges previous work published in the journal.

Wiley (2014):

Journals should facilitate academic debate. This involves encouraging correspondence and constructive criticism of the work the journals publish.

If an item of correspondence discusses a specific article, the journal should invite the authors of the work to respond before the correspondence is published. When possible, the correspondence and the authors' response should be published at the same time.

Rule 2: Errors together with inaccurate / misleading statements must be corrected promptly and with due prominence.

COPE (2011):

12.1 Errors, inaccurate or misleading statements must be corrected promptly and with due prominence

ICMJE (2022):

"Honest errors are a part of science and publishing and require publication of a correction when they are detected. Corrections are needed for errors of fact. Matters of debate are best handled as letters to the editor, as print or electronic correspondence, or as posts in a journal-sponsored online forum."

We self-derived our third rule from the three sources cited above.

Rule 3: Editors should behave with integrity, which is the quality of being honest and having strong moral principles.

We are of the opinion that the present and the past editor of PoPS broke these three rules for ethical behavior of journal editors in the present case. The remedy that we request is simply that APS allow us to publish our rebuttal.

3.4.1 APS' reply to our intial email

October 5, 2022

From: <Walker, Elaine>

To: <Bryan Pesta>

Sent: Wed, Oct 5, 2022 8:45 am

Subject: FW: [External] RE: Your submission to Perspectives on Psychological Science

Dear Dr. Pesta,

In my capacity as Chair of the APS Publications Committee, I am writing in response to your e-mail to Dr. Gopnick about the disposition of the article you submitted to the APS journal, *Perspectives on Psychological Science* (PoPS). I understand the basis for some of your concerns. It does indeed seem that, during the transition from the past to the current editor of the journal, some issues were not addressed in a timely manner and communication was not as efficient as it should have been.

The first point I want to address is an error made regarding the letter you received from Dr. Fiedler regarding your revised submission. Specifically, you did not receive the complete and more detailed letter from the editor that Dr. Fiedler had drafted and intended to send. As you are aware, Dr. Fiedler is a relatively new editor, and he and his staff were only in their positions for a short period of time when your revised submission was reviewed. I have attached to this email the full letter that Dr. Fiedler had intended to send. It is likely that a clerical oversight was responsible for your not receiving it. As Chair of the publications committee, I am sorry for the error.

I would also like to add that after reviewing all of the communications, I can see that some issues arose with the processing of the initial submission of your response to Giangrande and Turkheimer. The response time by Dr. King to your communications was too long. Also, the fact that the editor did not reach out to you and your coauthors about Giangrande and Turkheimer's article prior to its publication was undoubtedly a disappointment to you. Some other editors may indeed have done so. And some might even view it as a convention to do so. But in the majority of cases, the original article and the critique would be published in the same journal. In this case, it was the prerogative of the journal editor at the time, Dr. King, to refrain from making an invitation for a response to a critique of a paper that had been published in another journal. I reviewed the COPE guidelines on a post-publication critiques and I was not able to find any examples or advice for editors that involved procedures for processing critiques of papers published in another journal.

As you know, the demands on editors and reviewers have increased significantly as the number of journals and the volume of submissions have increased. Scientific organizations must, therefore, assure some autonomy for Editors and avoid burdening them with the responsibility for unpopular or contested decisions made by former editors. Also, when Editors process manuscripts that critique articles that were published in a journal other than their own, there are no precedents or guidelines governing the ethical obligations of the Editor.

In sum, despite the missteps along the way, it is my judgment that this situation you encountered did not involve any violation of the guidelines of the Committee on Publication Ethics' (COPE). I will be happy to provide additional clarification if needed.

Regards,

3.4.3 Follow-up reply to APS

October 14, 2022

Dear Dr. Walker,

Thank you for processing our complaint of unethical behavior against two PoPS editors, and for your detailed response. We appreciate your swift reply to our complaint. We also appreciate your offer to provide additional clarification if needed because we agree with you on specific topics, and we disagree on other topics. We discuss these issues below.

The decision is at the editor's discretion

The COPE guidelines state:

14.2 Authors of criticized material should be given the opportunity to respond

The COPE guidelines do not state:

It is at the discretion of the editor whether authors of criticized material should be given the opportunity to respond.

The COPE guidelines clearly state that we have a right to respond, so it is not only at the editor's discretion to allow criticized authors to respond.

Our original paper was published in *Intelligence* – Why is this relevant?

You write:

I reviewed the COPE guidelines on a post-publication critiques and I was not able to find any examples or advice for editors that involved procedures for processing critiques of papers published in another journal.

Pesta et al. (2020) published a paper in *Intelligence*. Suppose Giangrande and Turkheimer had submitted their critique of Pesta et al. to *Intelligence*, then it would have been almost a certainty that one of the five authors of Pesta et al. (2020) would be invited as a reviewer, so they would have to confront their opponents head-on. However, Giangrande and Turkheimer bypassed the review procedure of *Intelligence*, thereby reducing the chance they would have one of us as a reviewer. Instead, they published their critique in PoPS. Indeed, it is highly unusual and quite cowardly for scientists to evade a discussion with opponents. That is why you did not find any examples or advice for editors that involve procedures for processing critiques of papers published in another journal.

We would argue that your observation concerning "critiques of papers published in another journal" does not apply because the G&T paper was published in PoPS, and our critique of G&T's paper was submitted to PoPS rather than to *Intelligence*. You offered to provide additional clarification if needed, so would you be so kind as to explain why your point here is relevant?

No comment on our second complaint

The team that produced Pesta et al. (2020) consisted of five people who combined their considerable expertise. We show that Giangrande and Turkheimer's paper contains many inaccurate/misleading statements. To give just one of many possible examples: G&T make laughably silly comments on meta-analysis. The relevant COPE statement reads:

12.1 Errors, inaccurate of misleading statements must be corrected promptly and with due prominence

In your reply, we see no comment about our second complaint. You offered to provide additional clarification if needed, so we invite you to comment on this.

Dr. Fiedler incorrectly claimed he sent our paper out for review.

We would like to thank you for inviting Dr. Fiedler to give a detailed reason for why he rejected our paper. We have read his letter with great interest. Unfortunately, however, he still did not supply the feedback from various expert reviewers that he referred to in his rejection letter dated June 3rd, 2022, in which he writes:

Thank you for submitting your manuscript # PPS-22-163 entitled "On Group Differences in the Heritability of Intelligence: A Reply to Giangrande and Turkheimer (2022)" to Perspectives on Psychological Science. *I was happy to be offered the help of some of the most renowned expert reviewers for the evaluation of this manuscript. Please find their comments below. Based on the reviewers' advice* [Italics added] and on my own careful reading of your manuscript, I have concluded that this manuscript is not suitable for publication in Perspectives. I am therefore sorry to write that I have decided to decline the manuscript.

If these reviews existed, surely Dr. Fiedler would have supplied them to us by now. Moreover, Dr. Fiedler argued that it would not be a good idea to send out our revised manuscript for review. There is no doubt in our minds that these claimed reviews do not exist.

This intended letter indirectly makes it clear that Dr. Fiedler acted unethically by writing earlier that his rejection decision was partially based on renowned expert reviewers. Yet we have not received these reviews despite multiple requests by us.

At the end of your letter, you state that:

In sum, despite the missteps along the way, it is my judgment that this situation you encountered did not involve any violation of the guidelines of the Committee on Publication Ethics' (COPE).

You offered to provide additional clarification if needed, so would you be so kind as to discuss how presenting non-existing reviews from experts as a basis for a rejection of a paper is within the bounds of acceptable ethical behavior for editors?

Our offer to submit a much shorter, new manuscript that presents original evidence

In his Intended Decision Letter for PPS-22-163, PoPS editor Dr. Fiedler made various statements, which we quote:

… I cannot now publish a new article submitted under a different manuscript ID, which is however **nearly identical with the rejected manuscript**. The overlap is huge; the new manuscript represents nothing but a very light editing of the rejected one.

…

Indeed, I did write in my action letter "… that the door is open for your submitting a new manuscript that deals with the G&T critique in a convincing scientific style, based on original evidence or on an upfront discussion of methodological problems that are of interest for a broader readership, not just for a few personally involved authors."

…

However, for such a new manuscript to be considered, it has to present original evidence in a scientifically convincing style.

First, we agree with Dr. Fiedler that the original manuscript submitted to PoPS had substantial overlap with our revised version. Second, Dr. Fielder wrote on two occasions, "… that the door is open for your submitting a new manuscript …". We suggest here a way forward that will satisfy both of Dr. Fiedler's concerns.

Giangrande and Turkheimer generated a plethora of critiques against Pesta et al. (2020), but surprisingly, they never tested their hypotheses empirically. We offer now to fundamentally rewrite the paper and focus on a thorough, empirical, meta-analytical test of G&T's critiques. This new manuscript would satisfy two of Dr. Fiedler's requirements: 1) a severe reduction in length and 2) the use of original, empirical evidence. We then leave it up to Dr. Fiedler to send the manuscript to the original two reviewers, or to invite new reviewers; however, he should not desk-reject our new version.

In sum, we appreciate your offer to provide additional clarification and to engage in this ongoing discussion with you. We look forward to receiving your reply.

Sincerely,

References

Allen, G., & Pettigrew, K. D. (1973). Technical comment: Heritability of IQ by social class: Evidence inconclusive. *Science, 182*(4116), 1042–1047.

Baier, T., & Lang, V. (2019). The social stratification of environmental and genetic influences on education: New evidence using a register-based twin sample. *Sociological Science, 6*, 143–171.

Bares, C. B., Kendler, K. S., & Maes, H. H. (2016). Racial differences in heritability of cigarette smoking in adolescents and young adults. *Drug and Alcohol Dependence, 166*, 75–84.

Biagini, J. M., Kroner, J. W., Gonzales, A., He, H., Stevens, M., Grashel, B., … Hershey, G. K. K. (2022). Longitudinal atopic dermatitis endotypes: An atopic march paradigm that includes Black children. *Journal of Allergy and Clinical Immunology, 149*(5), 1702–1710.

Bronfenbrenner, U., & Ceci, S. J. (1994). Nature-nuture reconceptualized in developmental perspective: A bioecological model. *Psychological Review, 101*(4), 568.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

DeFries, J. C., Ashton, G. C., Johnson, R. C., Kuse, A. R., McClearn, G. E., Mi, M. P., … Wilson, J. R. (1976). Parent–offspring resemblance for specific cognitive abilities in two ethnic groups. *Nature, 261*(5556), 131–133.

Deutsch, M., Katz, I., & Jensen, A. R. (1968). In *Social class, race, and psychological development*. New York: Holt, Rinehart & Winston.

Engelhardt, L. E., Church, J. A., Paige Harden, K., & Tucker-Drob, E. M. (2019). Accounting for the shared environment in cognitive abilities and academic achievement with measured socioecological contexts. *Developmental Science, 22*(1), Article e12699.

Enkhmaa, B., Anuurad, E., Zhang, W., Kim, K., & Berglund, L. (2019). Heritability of apolipoprotein (a) traits in two-generational African-American and Caucasian families [S]. *Journal of Lipid Research, 60*(9), 1603–1609.

Fuerst, J., (2014). ACE Analysis of the NLSY79 AFQT by Race/Ethnicity. Human Varieties.

Fuerst, J., & Dalliard, M. (2014). *Genetic and environmental determinants of IQ in Black, White, and Hispanic Americans: A meta-analysis and new analysis*. Open Behavioral Genetics.

Giangrande, E. J., & Turkheimer, E. (2022). Race, ethnicity, and the Scarr-Rowe hypothesis: A cautionary example of fringe science entering the mainstream. *Perspectives on Psychological Science, 17*(3), 696–710.

Guo, G., & Stearns, E. (2002). The social influences on the realization of genetic potential for intellectual development. *Social Forces, 80*(3), 881–910.

Guo, G., & Wang, J. (2002). The mixed or multilevel model for behavior genetic analysis. *Behavior Genetics, 32*(1), 37–49.

Gusev, A., Shi, H., Kichaev, G., Pomerantz, M., Li, F., Long, H. W., … Pasaniuc, B. (2016). Atlas of prostate cancer heritability in European and African-American men pinpoints tissue-specific regulation. *Nature Communications, 7*(1), 1–13.

Haier, R. J. (2020). Academic freedom and social responsibility: Finding a balance. *Intelligence, 82*, Article 101482.

Halpern-Manners, A., Marahrens, H., Neiderhiser, J. M., Natsuaki, M. N., Shaw, D. S., Reiss, D., & Leve, L. D. (2020). The intergenerational transmission of early educational advantages: New results based on an adoption design. *Research in Social Stratification and Mobility, 67*, Article 100486.

Harden, K. P., Turkheimer, E., & Loehlin, J. C. (2007). Genotype by environment interaction in adolescents' cognitive aptitude. *Behavior Genetics, 37*, 273–283.

Hart, S. A., Soden, B., Johnson, W., Schatschneider, C., & Taylor, J. (2013). Expanding the environment: Gene× school-level SES interaction on reading comprehension. *Journal of Child Psychology and Psychiatry, 54*(10), 1047–1055.

Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Hodges, P., Juarez, A., & Gardner, M. (1976). *Estimates of heritability in different populations (A preliminary report)*.

Holden, L. R., Haughbrook, R., & Hart, S. A. (2021, December 22). Developmental behavioral genetics research on school achievement is missing vulnerable children, to our detriment. *PsyArXiv*. https://doi.org/10.31234/osf.io/gf3hc

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis*. London: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Jensen, A. R. (1968). Social class, race, and genetics: Implications for education. *American Educational Research Journal, 5*(1), 1–42.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Prager.

Khan, S. S., Kim, K. Y. A., Peng, J., Aguilar, F. G., Selvaraj, S., Martinez, E. E., … Shah, S. J. (2019). Clinical correlates and heritability of cardiac mechanics: The HyperGEN study. *International Journal of Cardiology, 274*, 208–213.

Kolifarhood, G., Daneshpour, M., Hadaegh, F., Sabour, S., Saadati, H. M., Haghdoust, A. A., … Khosravi, N. (2019). Heritability of blood pressure traits in diverse populations: A systematic review and meta-analysis. *Journal of Human Hypertension, 33*(11), 775–785.

Lee, J., & Zhou, M. (2015). *The Asian American achievement paradox*. Russell Sage Foundation.

Moeyaert, M., Ugille, M., Natasha Beretvas, S., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology*, 1–14.

Mollon, J., Knowles, E. E., Mathias, S. R., Gur, R., Peralta, J. M., Weiner, D. J., … Glahn, D. C. (2021). Genetic influence on cognitive development between childhood and adulthood. *Molecular Psychiatry, 26*(2), 656–665.

Munn-Chernoff, M. A., Grant, J. D., Agrawal, A., Koren, R., Glowinski, A. L., Bucholz, K. K., … Duncan, A. E. (2015). Are there common familial influences for major depressive disorder and an overeating–binge eating dimension in both European American and African American female twins? *International Journal of Eating Disorders, 48*(4), 375–382.

Musani, S. K., Martin, L. J., Woo, J. G., Olivier, M., Gurka, M. J., & DeBoer, M. D. (2017). Heritability of the severity of the metabolic syndrome in whites and blacks in 3 large cohorts. *Circulation. Cardiovascular Genetics, 10*(2), Article e001621.

Nichols, P. L. (1970). *The effects of heredity and environment on intelligence test performance in 4 and 7 year white and negro sibling pairs* (Doctoral dissertation, University of Minnesota).

Nielsen, F. (2016). The status-achievement process: Insights from genetics. *Frontiers in Sociology, 1*, 9.

Osborne, R. T. (1980). *Twins, black and white*. Foundation for Human Understanding.

Osborne, R. T., & Gregor, A. J. (1968). Racial differences in heritability estimates for tests of spatial ability. *Perceptual and Motor Skills, 27*(3), 735–739.

Osborne, R. T., & Miele, F. (1969). Racial differences in environmental influences on numerical ability as determined by heritability estimates. *Perceptual and Motor Skills, 28*(2), 535–538.

Pesta, B. J., Kirkegaard, E. O. W., te Nijenhuis, J., Lasker, J., & Fuerst, J. G. R. (2020). Racial and ethnic group differences in the heritability of intelligence: A systematic review and meta-analysis. *Intelligence, 78*, Article 101408.

Polubriaginof, F. C., Vanguri, R., Quinnies, K., Belbin, G. M., Yahi, A., Salmasian, H., … Tatonetti, N. P. (2018). Disease heritability inferred from familial relationships reported in medical records. *Cell, 173*(7), 1692–1704.

PRB. (2010). Shifting Latino ethnic and racial identity. PRB. https://www.prb.org/resources/shifting-latino-ethnic-and-racial-identity/

Rhea, S. A. (2015). Reviving the Louisville twin study: An introduction. *Behavior Genetics, 45*(6), 597–599.

Rhemtulla, M., & Tucker-Drob, E. M. (2012). Gene-by-socioeconomic status interaction on school readiness. *Behavior Genetics, 42*(4), 549–558.

Rindermann, H., Becker, D., & Coyle, T. R. (2020). Survey of expert opinion on intelligence: Intelligence research, experts' background, controversial issues, and the media. *Intelligence, 78*, Article 101406.

Rowe, D. C., Jacobson, K. C., & Van den Oord, E. J. (1999). Genetic and environmental influences on vocabulary IQ: Parental education level as moderator. *Child Development, 70*(5), 1151–1162.

Sartor, C. E., Nelson, E. C., Lynskey, M. T., Madden, P. A., Heath, A. C., & Bucholz, K. K. (2013). Are there differences between young African-American and European-American women in the relative influences of genetics versus environment on age at first drink and problem alcohol use? *Alcoholism: Clinical and Experimental Research, 37*(11), 1939–1946.

Scarr, S. (1981). Having the last word. In S. Scarr (Ed.), *Race, social class and individual differences in I.Q.: New studies of old issues* (pp. 261–315). Hillsdale, NJ: Erlbaum.

Scarr, S. (2009). Epilogue. In K. McCartney, & R. A Weinberg (Eds.), *Experience and development: A festschrift in honor of Sandra Wood Scarr* (pp. 253–268). New York: Psychology Press.

Scarr, S., & Barker, W. (1981). The effects of family background: A study of cognitive differences among black and white twins. In S. Scarr (Ed.), *Race, social class and individual differences in I.Q.: New studies of old issues* (pp. 261–315). Hillsdale, NJ: Erlbaum.

Scarr, S., & Weinberg, R. A. (1976). IQ test performance of black children adopted by white families. *American Psychologist, 31*(10), 726.

Scarr, S., Weinberg, R. A., & Waldman, I. D. (1993). IQ correlations in transracial adoptive families. *Intelligence, 17*(4), 541–555.

Scarr-Salapatek, S. (1971a). Race, social class, and IQ: Population differences in heritability of IQ scores were found for racial and social class groups. *Science, 174* (4016), 1285–1295.

Scarr-Salapatek, S. (1971b). Unknowns in the IQ equation. *Science, 174*, 1223–1228.

Scarr-Salapatek, S. (1973). Response: Heritability of IQ by social class: Evidence inconclusive. *Science, 182*(4116), 1045–1047.

Schmid, C. H., Stijnen, T., & White, I. (Eds.). (2020). *Handbook of meta-analysis*. CRC Press.

Schmidt, F., & Hunter, J. (2015). *Methods of meta-analysis* (Third Edition ed.). SAGE Publications, Ltd.

Schmidt, F., & Le, H. (2004). *Software for the Hunter-Schmidt meta-analysis methods* (p. 52242). Iowa City, IA: University of Iowa, Department of Management and Organizations.

Schwartz, J. A. (2015). Socioeconomic status as a moderator of the genetic and shared environmental influence on verbal IQ: A multilevel behavioral genetic approach. *Intelligence, 52*, 80–89.

Song, C., Peacor, S. D., Osenberg, C. W., & Bence, J. R. (2020). An assessment of statistical methods for nonindependent data in ecological meta-analyses. *Ecology, 101*(12), Article e03184.

Steinsaltz, D., Dahl, A., & Wachter, K. W. (2020). On negative heritability and negative estimates of heritability. *Genetics, 215*(2), 343–357.

Tucker-Drob, E. M., & Bates, T. C. (2016). Large cross-national differences in gene× socioeconomic status interaction on intelligence. *Psychological Science, 27*(2), 138–149.

Tucker-Drob, E. M., Rhemtulla, M., Harden, K. P., Turkheimer, E., & Fask, D. (2011). Emergence of a gene× socioeconomic status interaction on infant mental ability between 10 months and 2 years. *Psychological Science, 22*(1), 125–133.

Turkheimer, E. (1990). On the alleged independence of variance components and group differences. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition, 10*(6), 686–690.

Turkheimer, E., Beam, C. E., & Davis, D. W. (2015). The Scarr-Rowe interaction in complete seven-year WISC data from the Louisville twin study: Preliminary report. *Behavior Genetics, 45*(6), 635–639.

Turkheimer, E., Haley, A., Waldron, M., d'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14*(6), 623–628.

Turkheimer, E., Harden, K. P., D'Onofrio, B., & Gottesman, I. I. (2009). The Scarr–Rowe interaction between measured socioeconomic status and the heritability of cognitive ability. In K. McCartney, & R. A. Weinberg (Eds.), *Experience and development: A festschrift in honor of Sandra Wood Scarr* (pp. 81–97). New York: Psychology Press.

Turkheimer, E., Harden, K. P., & Nisbett, R. E. (2017, May 18). *Charles Murray is once again peddling junk science about race and IQ*. Vox.

U.S. Census. (2021). U.S. Census Bureau Guidance on the presentation and comparison of race and hispanic origin data Accessed at: https://www.census.gov/topics/population/hispanic-origin/about/comparing-race-and-hispanic-origin.html.

Uchikoshi, F., & Conley, D. (2021). Gene-environment interactions and school tracking during secondary education: Evidence from the US. *Research in Social Stratification and Mobility, 76*, Article 100628.

UK Government. (2019). Ethnicity facts and figures: List of ethnic groups Accessed at: https://www.ethnicity-facts-figures.service.gov.uk/ethnic-groups.

Van Den Oord, E. J., & Rowe, D. C. (1997). An examination of genotype-environment interactions for academic achievement in an U.S. national longitudinal survey. *Intelligence, 25*(3), 205–228.

Vandenberg, S. G. (1970). A comparison of heritability estimates of US Negro and White high school students. *Acta Geneticae Medicae et Gemellologiae, 19*(1–2), 280–284.

Warne, R. T. (2021). Between-group mean differences in intelligence in the United States are> 0% genetically caused: Five converging lines of evidence. *The American Journal of Psychology, 134*(4), 480–501.

Whiteman, M., & Deutsch, M. (1968). Social disadvantage as related to intellective and language development. In M. Deutsch, Katz, I., & A. R. Jensen (Eds.), *Social class, race, and psychological development* (pp. 86–114). New York: Holt, Rinehart & Winston.

Woodley of Menie, M. A, Figueredo, A. J., Dunkel, C. S., & Madison, G. (2015). Estimating the strength of genetic selection against heritable g in a sample of 3520 Americans, sourced from MIDUS II. *Personality and Individual Differences, 86*, 266–270.