


Challenging the Link Between Early Childhood Television Exposure and Later Attention Problems: A Multiverse Approach



Matthew T. McBee¹, Rebecca J. Brand², and
Wallace E. Dixon, Jr.¹ 

¹Department of Psychology, East Tennessee State University, and ²Department of Psychology, Villanova University

Psychological Science
1–23
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797620971650
www.psychologicalscience.org/PS


Abstract

In 2004, Christakis and colleagues published an article in which they claimed that early childhood television exposure causes later attention problems, a claim that continues to be frequently promoted by the popular media. Using the same National Longitudinal Survey of Youth 1979 data set ($N = 2,108$), we conducted two multiverse analyses to examine whether the finding reported by Christakis and colleagues was robust to different analytic choices. We evaluated 848 models, including logistic regression models, linear regression models, and two forms of propensity-score analysis. If the claim were true, we would expect most of the justifiable analyses to produce significant results in the predicted direction. However, only 166 models (19.6%) yielded a statistically significant relationship, and most of these employed questionable analytic choices. We concluded that these data do not provide compelling evidence of a harmful effect of TV exposure on attention.

Keywords

media, TV, ADHD, attention development, multiverse analysis, computational reproducibility, garden of forking paths, open data, open materials

Received 6/6/19; Revision accepted 9/19/20

Psychological science can have a broad and deep impact on human lives. In development in particular, there is a sense of relevance, indeed urgency, to many of its questions: What are the causes of autism? Is it helpful or harmful to grow up multilingual? Does screen time cause attention deficits? The stakes are high; it is crucial that scientists get the answers right. Unfortunately, the replication crisis in the social and behavioral sciences has shown that many claims in the literature do not hold up to reexamination (Camerer et al., 2018; Open Science Collaboration, 2015).

Once an erroneous finding has been disseminated, it seems nearly impossible to correct public misconceptions. One salient example involves Andrew Wakefield's fraudulent claim (Wakefield et al., 1998, retracted) of a link between autism and the measles-mumps-rubella vaccine (Committee to Review Adverse Effects of Vaccines,

2012; Oliver & Wood, 2014). Whether due to fraud, mismanagement, or merely chance, nonreplicable findings derail scientific progress. Engaging in replication attempts and reanalyzing the robustness of reported findings are among the important strategies available for combating this replicability crisis by determining which claims hold up to increased scrutiny and reexamination (*Nature*, 2016).

In this article, we reexamine the study by Christakis et al. (2004), who claimed that there was a positive association between television exposure in toddlerhood and attention problems at school age. Although the

Corresponding Author:

Wallace E. Dixon, Jr., East Tennessee State University, Department of Psychology
E-mail: dixonw@mail.etsu.edu

study was longitudinal in nature and included a variety of control variables, its lack of randomized manipulation of TV use made it difficult to draw strong causal conclusions from the data. In our view, the provisional nature of Christakis et al.'s claim was carefully described in the article itself. However, less qualification was used in the lead author's subsequent public statements. For example, in a TEDx talk that has been viewed more than half a million times, the finding from this article was cited as evidence supporting the "overstimulation hypothesis," according to which "prolonged exposure to this rapid image change [from television] during this critical window of brain development . . . precondition[s] the mind to expect high levels of input and . . . lead[s] to inattention later in life" (Christakis, 2011, 6:36–6:53). He went on to say,

And we tested this some years ago, and what we found was that . . . the more television children watched before age three, the more likely they were to actually have attentional problems at school age. Specifically, for each hour that they watched before the age of three, their chances of having attentional problems was increased by about ten percent. So a child who watched two hours of TV a day before age three would be twenty percent more likely to have attention problems compared to a child who watched none. (Christakis, 2011, 7:19–7:46)

Three things are notable (and potentially falsifiable) about this claim: first, that the association actually exists; second, that it is causal (that TV exposure leads to later attention problems); and third, that the association is linear (for each unit of television exposure, one can predict a specific and constant increase in the probability of attention problems). If we are going to base policy and parenting guidance on the claim, we think it is important to confirm whether it is really true.

Subsequent research justifies skepticism regarding the claim. A reanalysis of the data set used by Christakis et al. (2004) indicated that the finding was not robust to certain small changes in model specification (Foster & Watkins, 2010). A recent meta-analysis on screen-media use and attention problems indicated not only that the relationship between them was, at best, a small to moderate one but also that even the direction of effect was unclear (Nikkelen et al., 2014; see also Kostyrka-Allchorne et al., 2017).

Given these more nuanced and updated findings, one might question whether a 17-year-old claim is worth further examination. However, the meme regarding the harmfulness of screen time is still deeply embedded in popular understanding. Using Google search in June 2020 for "does TV cause attention

Statement of Relevance

Psychological science can have a broad and deep impact on human lives, especially when its messages bear on the healthy development of children, when there is often a sense of urgency to many of its questions. It is essential that scientists continually work to correct inaccurate messages that have made it into public consciousness. One salient example is research that seemingly demonstrated that TV exposure in toddlerhood contributes to attention problems in school-age children. In the present research, we employed a multiverse analysis to re-examine this claim. After evaluating 848 statistical models of the possible association between TV exposure and attention problems, we concluded that the data do not provide compelling evidence of a harmful effect. We suspect the original result that TV was harmful probably emerged because of chance peculiarities found in the data, given that a similar outcome was not found in the vast majority of our analytical models.

problems," most of the top hits, including some from reputable sites such as WebMD and why.org, claim a link between TV and attention problems. WebMD uses blatantly causal language in its headline ("Toddler TV Time Can Cause Attention Problems"; Peck, 2004) and another site quotes Christakis as saying "TV 'rewires' an infant's brain" to cause attention deficit/hyperactivity disorder (ADHD; Lotus, 2020). It is also telling that the original article suggesting a link between TV exposure and attention problems was cited 118 times in a recent 2-year period (January 2017 to December 2018); during the same time frame, the more methodologically sound critique (Foster & Watkins, 2010) had 18 citations, and the meta-analysis (Nikkelen et al., 2014) had only 38.

Our goal in the current study was to examine the robustness of the original claim through use of a multiverse analysis (Silberzahn et al., 2018; Steegen et al., 2016; see also Orben et al., 2019). In any research endeavor, a series of analytic decisions must be made, some of them arbitrary or nearly so (King & Zeng, 2007). This series of decisions has been called the "garden of forking paths" (Gelman & Loken, 2013). If different paths through the garden lead to substantively different conclusions, a finding cannot be considered robust. One way to evaluate the dependence of a claim on a specific model is to subject the data to a wide variety of defensible analyses, systematically exploring how sensitive the outcome is to different model specifications.

In this article, we present two multiverse analyses of Christakis et al.'s (2004) original claim, using the same

Table 1. Conceptual Map of Analytic Decisions for the Multiverse Analyses

Feature	Levels
Multiverse I: logistic regression (504 total analyses, 200 nonredundant)	
Attention cut point	110, 112, 114, . . . 130 (21 levels)
Outcome variable	Within-sex standardized raw attention
Treatment of missing data	Listwise deletion versus multiple imputation
Sample weights	Incorporated versus not incorporated
Covariate set	Original versus expanded
TV exposure age	~1.5 years versus ~3 years
Multiverse II: linear regression (24 total analyses)	
Outcome variable	Within-sex standardized versus raw attention
Treatment of missing data	Listwise deletion versus multiple imputation
Sample weights	Incorporated versus not incorporated
Covariate set	Original versus expanded
TV exposure age	~1.5 years versus ~3 years
Propensity-score analysis, inverse-probability-of-treatment weighting (384 total analyses)	
Outcome variable	Within-sex standardized versus raw attention
TV-exposure cut point	20/80, 30/70, 40/60, 50, 60, 70 (percentiles, six levels)
Sample weights	Incorporated versus not incorporated
Covariate set	Original versus expanded
TV exposure age	~1.5 years versus ~3 years
Treatment effect	Average treatment effect for the treated versus average treatment effect
Doubly robust estimation	Doubly robust versus no additional covariate adjustment
Propensity-score analysis, stratification (240 total analyses)	
Outcome variable	Within-sex standardized versus raw attention
TV-exposure cut point	20/80, 30/70, 40/60, 50, 60, 70 (percentiles, six levels)
Sample weights	Incorporated versus not incorporated
Covariate set	Original versus expanded
TV exposure age	~1.5 years versus ~3 years
Number of strata	4, 5, 6, 7, 8 (five levels)

Note: Missing data in the propensity-score models were treated as informative by the boosted-classification-trees method used to estimate the propensity scores.

data set. The first employed logistic regression in conformance with the original study. The second expanded the range of approaches to include linear regression and propensity-score analysis (PSA) techniques. See Table 1 for a conceptual map of these analyses.

The first multiverse analysis closely corresponded with the original analysis, using logistic regression to predict school-age attention problems from toddler TV exposure. Following Christakis et al., we operationalized attention problems using the five-item subscale of the Behavior Problems Index (BPI) reflecting hyperactivity. These scores were nationally normed and are presented on an IQ-like metric, standardized within age and sex. We question the wisdom of using logistic regression, as it requires dichotomizing the continuous-attention variable with little justification for doing so. However, we conducted these analyses to explore whether we could replicate the original finding and to examine the impact of modeling and data-preparation

decisions within this framework. Because there is no apparent reason to prefer standardized scores in models that control for child sex and age, we also performed analyses using unstandardized (raw) attention scores by using the mean of the five relevant items.

In addition to deciding whether to dichotomize the outcome variable, we found the question of how to do so to be ambiguous. On a continuous measure of attention problems, where is the line between nonproblematic and problematic levels? With no a priori method in the literature for determining the appropriate cut point, Christakis et al. chose a score of 120,¹ arguing that this yielded a rate of problematic attention resembling the prevalence of ADHD in the population. In order to examine the sensitivity of the findings to this choice, we systematically varied the cut point from 110 to 130 on the standardized outcome and used percentile-equivalent cut points on the raw outcome. Finally, we also varied how to treat missing data in the logistic

regression—using listwise deletion, as in the original, or multiple imputation—and whether to use sample weights. Sample weights were used in the original but are discouraged by the National Longitudinal Survey of Youth (NLSY) in the case of regression and related modeling.² We ran models both ways.

The second multiverse analysis employed linear regression and two forms of PSA. Linear regression, unlike logistic regression, does not require the attention variable to be dichotomized and thus allowed us to directly test the claim that for each additional hour of TV exposure, we would see a rise in attention problems. Linear regression models are also a good choice because they are relatively high in statistical power and efficiency. However, the performance of such models is based on assumptions that can be difficult to justify and evaluate.

With the current data set, we believed that PSA was the most defensible choice for estimating a causal effect. Propensity scores are each case's predicted probability of being in the treatment group (in this case, the group being shown a large amount of TV), conditional on a variety of baseline characteristics (such as mother's education and household income; cf. Austin, 2011). Propensity scores are applied to produce virtual comparison groups that are balanced on all the covariates in the model. PSA thus approximates random assignment of subjects to two groups, helping to isolate TV as the potential causal variable. Relative to linear regression, PSA is more robust and less prone to systematic bias caused by violating model assumptions.

Within the propensity-score family of analyses, there are still many potential routes through the garden of forking paths. We explored a number of these branching routes (see Table 1):

Techniques: There are several techniques for implementing PSA. We selected two: inverse-probability-of-treatment weighting (IPTW) and stratification (Guo & Fraser, 2015). In stratification models, we varied the number of strata created from four to eight.

Outcome variable: As described above, we used both standardized and raw attention scores.

Attention cut point: Logistic models required the response variable to be divided into binary groups. We employed a range of 21 different cut points for making this division.

TV-exposure cut point: As a technique for virtual experimentation, PSA requires dichotomizing the treatment variable into something like a treatment group and a control group. Although dichotomization

of a continuous variable is justified here because of the advantages of PSA, understanding how the cut point influences the outcome would be informative. If the effect of TV on attention is linear across the range, then the effect should be proportional to the difference in median TV exposure between the groups regardless of where the cut point is situated. To explore this question, we ran analyses using six different cut points.

Sample weights: In IPTW propensity-score, linear regression, and logistic models, we ran analyses with and without sample weights applied.

Covariate set: We conducted each analysis with two sets of covariates: the first exactly as in the original and the second with a superordinate set of theoretically motivated covariates (as detailed in the next section).

TV-exposure age: Following Christakis et al., we included TV exposure measured at two ages (~1.5 and ~3 years).

Treatment effect: Because an effect of TV exposure might be different for children who watch a lot and for the average child, in IPTW models, we calculated estimates for both the average treatment effect (ATE) and the average treatment effect for the treated (ATT). Stratification, linear regression, and logistic models estimated the ATE.

Doubly robust estimation: In each IPTW propensity-score model, we identified the four covariates with the largest residual imbalance statistics and gave those covariates an additional regression adjustment. This is referred to as a *doubly robust strategy* that offers protection against any remaining bias due to residual imbalance on the covariates after applying the propensity scores (Guo & Fraser, 2015).

In summary, we evaluated the claim that early childhood TV exposure is associated with increased mid-childhood attention problems. Using R (R Core Team, 2020), we conducted two multiverse analyses, using the same National Longitudinal Survey of Youth 1979 (NLSY79) data set, prepared in the same manner as in Christakis et al.'s (2004) article. We employed variations of logistic regression, as per the original study, and added linear regression and propensity-score models. In all, we examined the relation in 848 distinct ways. Evaluating the outcomes across hundreds of models allowed us to assess the robustness of the original claim and better understand how outcomes may be impacted by specific analytic choices. If the claim is true, we would expect most of the justifiable analyses to produce significant results in the predicted direction.

Method

Data

Following Christakis et al. (2004), we obtained data for the present investigation from the NLSY79, available via the NLS Investigator Web interface (<https://www.nlsinfo.org/investigator/pages/search?s=NLSY79>). We downloaded 340 variables from the NLSY79 Child and Young Adult data set and 40 variables from the original adult NLSY79 data set. See <https://osf.io/4u69g/> for access to the complete data set, analysis packages and code, and directions for reproducing our analysis in Docker software.

Our variable-selection process was based on the one reported in the original article. Following Christakis et al., we selected three cohorts of children who were approximately 7 years old during the three index years of 1996, 1998, and 2000. Our baseline variable selections conformed to those in the original study as closely as possible, given the text of the original article, which did not report ID codes for the selected variables. In most cases, we could unambiguously identify variables by searching the NLSY data by question text or question title.

Selection of cases. We followed the original authors' criteria for sample selection and subject exclusion. For each index year (1996, 1998, and 2000), we included those children whose ages at index were between 6 years 9 months and 8 years 9 months. In accordance with the original study, we excluded children with severe vision or hearing impairment, as well as those with severe emotional disturbances or orthopedic disabilities. We extracted a total of 2,108 cases that met these conditions.

Variables. As in the original study, our measure of attention was the standardized score on the hyperactivity subscale of the five-item BPI, which was standardized to an IQ-like metric ($M = 100$, $SD = 15$) within sex, as per the original study, which we hereafter refer to as the *within-sex standardized attention score*. However, we also retained the raw attention scores that were unadjusted for sex. The five BPI items addressed children's ability to concentrate and pay attention, as well as their confusion, impulsivity, obsessions, and restlessness or inability to sit still.

Television use was calculated as in the original study. Items measuring hours per day of television watched by the child on both weekdays and weekend days were converted to average hours of TV by multiplying weekday hours per day by 5, adding to this weekend hours per day multiplied by 2, and dividing by 7. We took this measurement from three and two waves prior to the index year, so that TV exposure was measured when children were approximately 1.5 and 3 years old,

though the exact age of each child during these waves could vary to some extent.

It was necessary to correct some out-of-range values prior to analysis. We followed the procedure described in the original article, truncating any out-of-range values of the following variables to the top of their ranges: TV exposure in average hours per day exceeding 16 to 16 and highest grade completed exceeding 24 to 24 (as this would imply more than 8 years of postgraduate education). One high value for annual income (\$839,078) was set to "missing" because a comment in the NLSY codebook indicates that this value is unreliable.

For a file listing how our substantive, conceptual variable names relate to the NLSY alphanumeric variable names, see "variable name propagation spreadsheet .xlsx" at <https://osf.io/4u69g/> (under the Documentation folder). The analysis code is the canonical description of how the variables were constructed and should resolve any vagueness or ambiguity in the preceding description.

Selection of covariates. The goal of each of our models was to estimate the causal effect of early TV on mid-childhood attention as accurately as possible. Because these data were collected via an observational longitudinal design, confounding was a serious concern. Causal inference from observational data, in theory, is possible if the proper set of covariates is incorporated into the analysis such that all confounding paths are blocked (Rohrer, 2018). To this end, our models employed two different sets of covariates.

Original covariates. The first set of covariates was identical to that employed in the original study. It consisted of the following: cohort (the year in which the child's attention was assessed: 1996, 1998, or 2000), the child's age when attention was assessed (typically 93 months, but it varied between 81 and 105 months), child's race, child's sex, the number of children of the mother living in the household, the mother's highest grade completed, the cognitive stimulation and emotional support of the home (measured between the ages 1 and 3 years), binary indicators of maternal alcohol use and cigarette smoking during pregnancy, a binary indicator of whether the child's father lived in the household, maternal self-esteem (as assessed in 1987 using the Rosenberg Self-Esteem Scale), maternal depression (as measured in 1992 using the Center for Epidemiologic Studies-Depression [CES-D] scale), child's gestational age at birth (centered at term), and an urbanicity indicator variable in the form of the four levels of the standard metropolitan statistical area (SMSA) classification. Where applicable, all of these were extracted from the first wave of data availability to avoid conditioning on posttreatment variables because

such variables could have potentially biased our estimates if they were mediators or colliders (Montgomery et al., 2018; Rohrer, 2018).

Expanded covariates. The expanded covariate list included all the original covariates with the following additions, which we suspected to be confounders for TV exposure and childhood attention. We added family income, the partner's or spouse's highest level of educational attainment, an indicator variable for low birth weight (less than 2,500 g, or 5 lb 8 oz), child temperament, and an indicator that the child suffered from a health condition that limited school and play activities.³ Rather than using a continuous gestational-age-at-birth variable, we created a binary indicator of preterm delivery (birth before 37 weeks of gestation), as we suspected this would better capture the relevant information in this variable.

Most variables were based on survey questions that were repeatedly administered on a biennial basis and were selected from survey administrations contemporaneous with the TV-exposure observation. However, two exceptions were maternal self-esteem, which was asked only in 1987, and maternal depression, which was assessed (using the CES-D) only in 1992. Depending on the cohort, depression could have been assessed up to 4 years before or the same year the child was born, and self-esteem could have been assessed from 1 to 5 years before the child's birth. Despite this problem of timing, we included these two variables because the original authors did. However, we also expected a moderate degree of stability over time in these constructs (Lovibond, 1998; Trzesniewski et al., 2003), which may ameliorate some concern about the timing of their measurement. We hoped that including these covariates would reduce the confounding bias that would otherwise render the estimates uninterpretable, though it is unlikely that we eliminated it entirely (Westfall & Yarkoni, 2016).

One of our added covariates was child's temperament. Temperament includes the ability to regulate one's own attention (Posner & Rothbart, 2018; Smith et al., 1997), and as one might predict, certain temperament dimensions predict children's later attention problems (Auerbach et al., 2008; Sullivan et al., 2015). In addition, parents' perception of infants' energy level (Nabi & Krcmar, 2016), poor self-regulation (Radesky et al., 2014), and fussiness (Thompson et al., 2013) all predict TV exposure, suggesting that parents may be showing TV to infants as a way to manage their difficult temperaments. In short, we suspected that relations between early TV exposure and later attention problems, to the extent that they exist, might be driven by their shared connection to early attention problems (as reflected in temperament).

Our temperament scale was based on the temperament items included in the NLSY data set (National Longitudinal Surveys, 2020). We summed the six available items that represented aspects of difficult temperament, as defined by Rothbart and Bates (2006), which included irritability, high-intensity affect, and negative mood. These items included assessments of how often the child cries when seeing a stranger, how often the child is afraid of dogs or cats, how often the child cries with doctors or nurses, how often the caregiver has trouble calming the child, and how often the child cries compared with others. Our temperament variable was the mean of these items, each of which was represented on a 5-point scale.

Because reviewers expressed concern that our temperament items might simply reflect attention deficits assessed earlier in life, we performed an exploratory factor analysis of the temperament and attention items. A two-factor model with varimax rotation exhibited clean simple structure separating attention from temperament, and the largest absolute standardized cross-loading was .133. The correlation (r) between factors was $-.114$, 95% confidence interval (CI) = $[-.039, .185]$. We therefore concluded that attention and temperament were sufficiently distinct variables.

Analytic approaches

For each of the following analytic approaches, we modeled two different outcomes (raw attention vs. the within-sex standardized attention scores used in the original analysis), measured TV exposure at approximately 1.5 and 3 years of age, and incorporated the two different sets of covariates designated above. Additional features specific to each model are described below and in Table 1.

Multiverse 1 (logistic regression). All analyses were performed using R (Version 3.6.3; R Core Team, 2020). First, to replicate the analysis used in the original study, we analyzed the data set using logistic regression. As already noted, Christakis et al. divided the continuous-attention/behavior-problems scale into typical and problematic levels of attention on the basis of a standardized attention cut point of 120. To determine how sensitive the original findings were to this particular cut point, we defined multiple dichotomous outcome variables by varying the standardized attention cut point from 110 to 130. For comparison between analyses using the raw versus the standardized attention measure, we used cut points on the raw attention measure that were the percentile equivalents of those on the standardized attention measure.

We fitted models both with and without sample weights, using the *survey* package (Version 4.0; Lumley,

2014) to perform the weighted analysis. We also fitted models both with and without multiple imputation of missing data, using the *mice* package (Version 3.8.0; van Buuren & Groothuis-Oudshoorn, 2011). However, it was not possible to fit models using both sample weights and multiple imputation simultaneously. Listwise deletion yielded 336 models—21 (attention cut point) \times 2 (outcome) \times 2 (TV-exposure age) \times 2 (covariate set) \times 2 (sample weight)—and multiple imputation yielded 168 models—21 (attention cut point) \times 2 (outcome) \times 2 (TV-exposure age) \times 2 (covariate set)—for a total of 504 logistic regression models. However, because of sparseness on the attention outcome (particularly in the raw version), the imposition of two adjacent cutoffs (e.g., 121 and 122) would frequently produce identical categorizations of the outcome and therefore redundant results. After purging these redundancies, we were left with 200 unique logistic regression models.

Multiverse II: linear regression and PSA.

Linear regression. These models estimated the linear relationship between TV exposure, measured at the approximate ages of 1.5 and 3 years, and the midchildhood standardized and raw attention outcomes. They were the only models that treated both TV and attention as continuous variables. As with the logistic regressions, we fitted models both with and without sample weights and with and without multiple imputation (using the *survey* and *mice* packages, respectively). Again, sample weights could not be combined with multiple imputation, so these conditions were not fully crossed. Using listwise deletion, we fitted 16 models—2 (outcome) \times 2 (TV-exposure age) \times 2 (covariate set) \times 2 (sample weight)—whereas using multiple imputation, we fitted eight models—2 (outcome) \times 2 (TV-exposure age) \times 2 (covariate set)—for a total of 24 linear regression models.

Propensity-score analyses. Finally, we conducted PSAs using two techniques for incorporating the propensity scores (IPTW vs. stratification). We ran analyses using both the raw and within-sex standardized versions of the outcome at the approximate ages of 1.5 years and 3 years. To explore the impact of how hours of TV are dichotomized into high and low groups, we ran analyses using six different percentile cut points: with cutoffs at the 50th, 60th, and 70th percentiles, as well as 20/80 (i.e., below 20th percentile/above 80th percentile), 30/70, and 40/60.

Where possible, we ran analyses with and without a doubly robust strategy, with and without sample weights, and for both the ATT and the ATE. In all the PSAs, we used boosted classification trees (as implemented in the *twang* package, Version 1.6, Ridgeway et al., 2017) to estimate the propensity scores, using

bagging and cross-validation to prevent overfitting. Missing data on covariates are handled automatically by the classification-tree approach, in that the missingness is treated as informative, and propensity scores can be estimated for cases with missing covariate values.

Inverse-probability-of-treatment weighting. Using IPTW, we were able to fully cross all conditions, yielding 384 IPTW propensity-score models—6 (TV-exposure cut point) \times 2 (outcome: raw vs. standardized) \times 2 (TV-exposure age: 1.5 vs. 3) \times 2 (covariate set) \times 2 (treatment effect: ATT vs. ATE) \times 2 (sample weight) \times 2 (doubly robust vs. not doubly robust). The *survey* package (Version 4.0; Lumley, 2014) was used to estimate the treatment effect after applying IPTW.

Stratification. Two hundred forty stratification propensity-score models were computed—five different numbers of strata (4, 5, 6, 7, or 8) were fully crossed with the variables TV-exposure cut point (six levels), outcome (two levels), TV-exposure age (two levels), and covariate set (two levels). Neither sample weights nor the doubly robust approach could be implemented in the stratification models, nor could these models estimate the ATT. We used the *PSAgraphics* package (Version 2.1.1; Helmreich & Pruzek, 2009) to perform the stratified analysis, and we calculated *p* values for the treatment-effect estimates using the normal approximation.

In total, we fitted 848 nonredundant models to the data, including 200 logistic regression models, 24 linear regression models, and 624 propensity-score models.

Results

Results are summarized here; see <https://osf.io/4u69g> for additional details on each model.

Descriptive statistics

Tables 2 and 3 provide descriptive statistics for the continuous and categorical variables, respectively. The scatterplots in Figure 1 illustrate the relationship between early TV exposure (at ~1.5 and ~3 years) and later attention measured at age ~7 years (standardized within sex). The top row of Figure 1 shows the relation without covariates, and the bottom row shows the relation after removing the influence of covariates. Because missing data on the covariates dramatically reduced the sample size for the available analyses, we display the imputed data taken from the first (of 10) multiple imputations using red “x” symbols in the bottom row. Figure 1 also contains nonparametric smoothed regression lines to help illustrate the relationship between TV exposure and attention. The solid blue line fits to complete

Table 2. Marginal Descriptive Statistics for Continuous Variables

Variable	Valid <i>n</i>	<i>M</i>	<i>SD</i>	Minimum	Maximum
Age (years) when attention was measured	2,108	7.75	0.61	6.75	8.75
Annual family income (thousands)	1,958	33.42	24.53	0.00	189.92
Attention (raw)	2,108	2.64	0.39	1.00	3.00
Attention within-sex (standardized)	2,075	101.25	13.79	83.00	136.00
CES-D score (assessed in 1992)	2,089	46.97	7.87	32.30	79.90
Cognitive stimulation of home at age 1–3 years	1,907	97.61	16.15	11.10	148.20
Emotional support of home at age 1–3 years	1,765	97.99	16.58	31.60	124.70
Gestational age at birth	1,960	–1.41	1.96	–14.00	7.00
Mother's age (years) at birth	2,108	28.48	2.62	23.00	36.00
Mother's years of schooling	2,095	12.95	2.48	0.00	20.00
Number of children in household	2,097	1.64	1.20	0.00	7.00
Partner's years of schooling	1,757	13.28	2.70	1.00	20.00
Rosenberg self-esteem score (assessed in 1987)	2,040	45.07	8.40	23.50	59.70
Temperament	1,961	2.01	0.69	1.00	5.00
TV hours per day at age 1.5 years	1,993	2.23	3.07	0.00	16.00
TV hours per day at age 3 years	2,023	3.68	3.12	0.00	16.00

Note: CES-D = Center for Epidemiologic Studies-Depression.

(nonimputed) data only, whereas the dashed red line fits to all the data, including the imputed portion.

Visual consideration of these scatterplots indicates an apparent lack of linear relationship between TV exposure and attention. The only relation evident from the smoothed trajectory is a slight nonlinear “wobble” in the 2- to 6-hr-per-day range of TV exposure. This nonlinearity is diminished but not eliminated by controlling for covariates but is almost completely absent under the imputation of missing data. This suggests that any association between TV exposure and attention could represent a combination of confounding- and missing-data bias.

Multiverse I results (logistic models)

For each analysis, we report the odds ratio (*OR*) of the relationship between TV exposure at the ages of approximately 1.5 and 3 years and the probability of a child being in the problematic category of attention after we controlled for covariates. *ORs* greater than 1 indicate a higher risk of being classified into the “problematic-attention” category. We varied the threshold for problematic attention from 110 to 130. Results are summarized in Figure 2. Effect-size point estimates and 95% confidence intervals are given in *OR* units.

The median *OR* was 1.036, with 1st and 3rd quartiles of [1.011, 1.072] and a median *p* value across all models of .213. Overall, 61 of 200 models (30.5%) produced significant estimates in the predicted direction, and none produced significant estimates in the opposite direction.

As shown Figure 2, the results of the logistic regression analysis were highly sensitive to the choice of

attention cut point. Statistically significant estimates of the relation between TV exposure measured at the age of about 3 years and the probability of attention problems began to appear at standardized attention cut points of 115 or above. For TV exposure measured at the age of about 1.5 years, significant relations with the probability of attention problems began to appear at 120 and above.

Results were also dependent on choices regarding sample weights and handling of missing data. As shown in Figure 3, statistical significance occurred at a higher rate when sample weights were applied (31/64, or 48.4%) than when they were not (30/136, or 22.1%). When considering only those models with no sample weights, we found that a higher proportion yielded significance under listwise deletion (18/64, or 28.1%) than under multiple imputation (12/72, or 16.7%). Further, this small percentage of significant estimates under multiple imputation and no sample weights—the conditions that we found most defensible—tended to be barely significant, as illustrated by the lower CI boundaries that nearly include 1. The median *p* value for these 12 significant tests was .034, and their median *OR* was 1.060.

We note that we did not exactly replicate the values reported by Christakis et al. under putatively identical models. Using the standardized attention outcome with a 120 cutoff, the original covariate set, listwise deletion, and sample weights, we estimated an *OR* of 1.137, 95% CI = [1.066, 1.214], *p* < .001 when TV exposure was measured at approximately 3 years; Christakis et al. reported an *OR* of 1.09, 95% CI = [1.03, 1.15] for this condition. We estimated an *OR* of 1.058, 95% CI = [0.987, 1.134], *p* = .114, when TV exposure was measured at

Table 3. Marginal Descriptive Statistics for Categorical Variables

Variable and value	<i>n</i>	Percentage of <i>N</i>
Maternal alcohol use in pregnancy		
No	1,050	49.81
Yes	932	44.21
Missing	126	5.98
Cohort (interview wave when attention was assessed)		
1996	829	39.33
1998	796	37.76
2000	483	22.91
Father absent from household		
No	1,681	79.74
Yes	399	18.93
Missing	28	1.33
Child sex		
Female	1,034	49.05
Male	1,074	50.95
Low birth weight (< 2,500 g)		
No	1,812	85.96
Yes	138	6.55
Missing	158	7.50
Health condition that limits school or play		
No	1,917	90.94
Yes	122	5.79
Missing	69	3.27
Premature birth		
No	1,744	82.73
Yes	216	10.25
Missing	148	7.02
Child race		
Black	572	27.13
Hispanic	397	18.83
White	1,139	54.03
Maternal smoking in pregnancy		
No	1,447	68.64
Yes	528	25.05
Missing	133	6.31
SMSA (urbanicity)		
Not in SMSA	382	18.12
SMSA; central city unknown	680	32.26
SMSA; in central city	302	14.33
SMSA; not central city	639	30.31
Missing	105	4.98

Note: SMSA = standard metropolitan statistical area.

approximately 1.5 years, whereas the original authors reported an *OR* of 1.09, 95% *CI* = [1.02, 1.16]. We cannot explain these discrepancies.

Multiverse II results

Linear regression models. Regression coefficients for the effect of TV on attention were standardized such that

the estimates represent the expected change in attention (in *SD* units) given a 1-hr change in TV exposure. The direction of the raw attention outcome has been reversed to be consistent with the standardized outcome; higher scores represent worse attention for both. Results are summarized in Figure 4. The median estimate (β) for these models was -0.002 with 1st and 3rd quartiles of $[-0.008, 0.013]$, median $p = .335$.

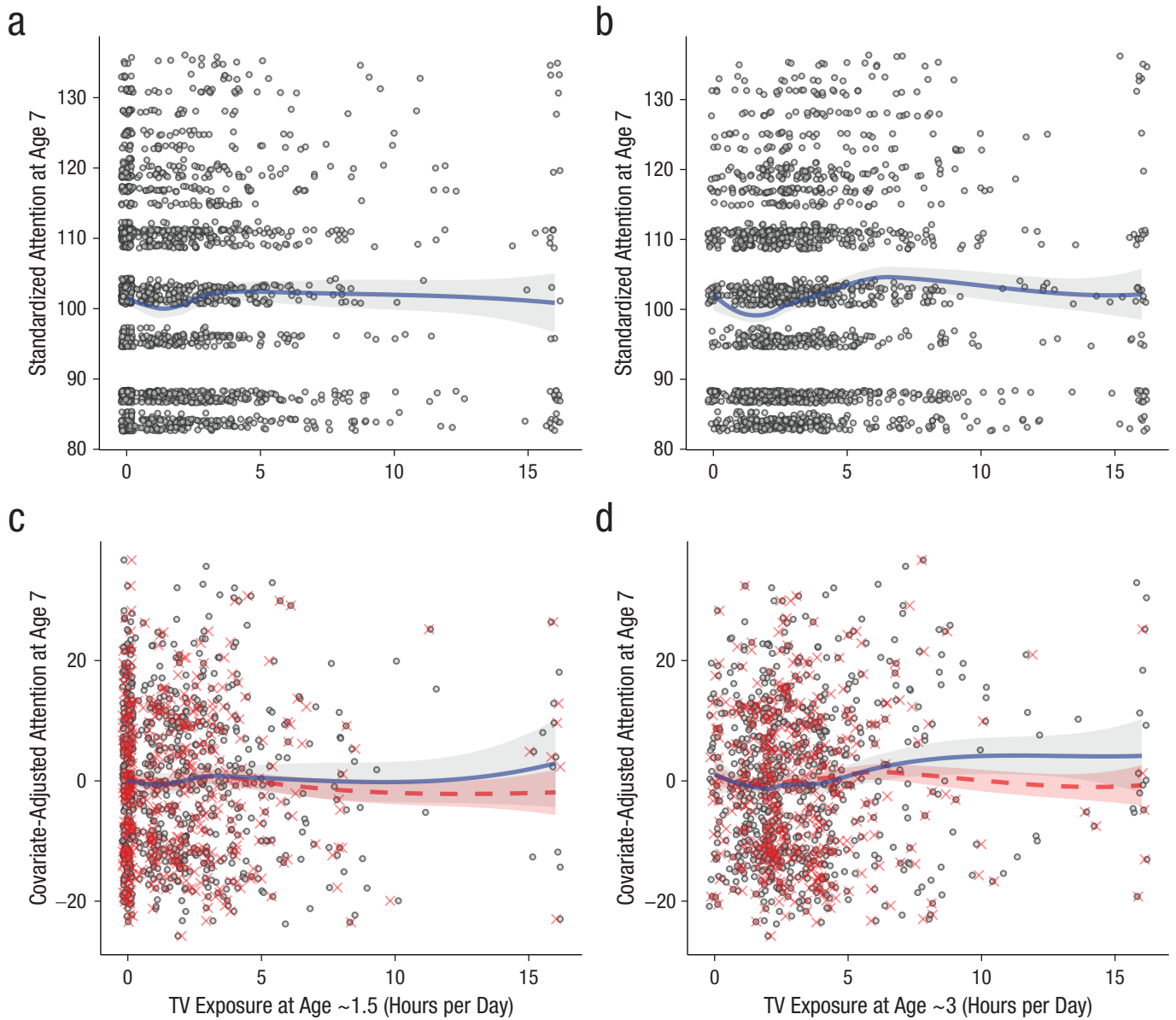


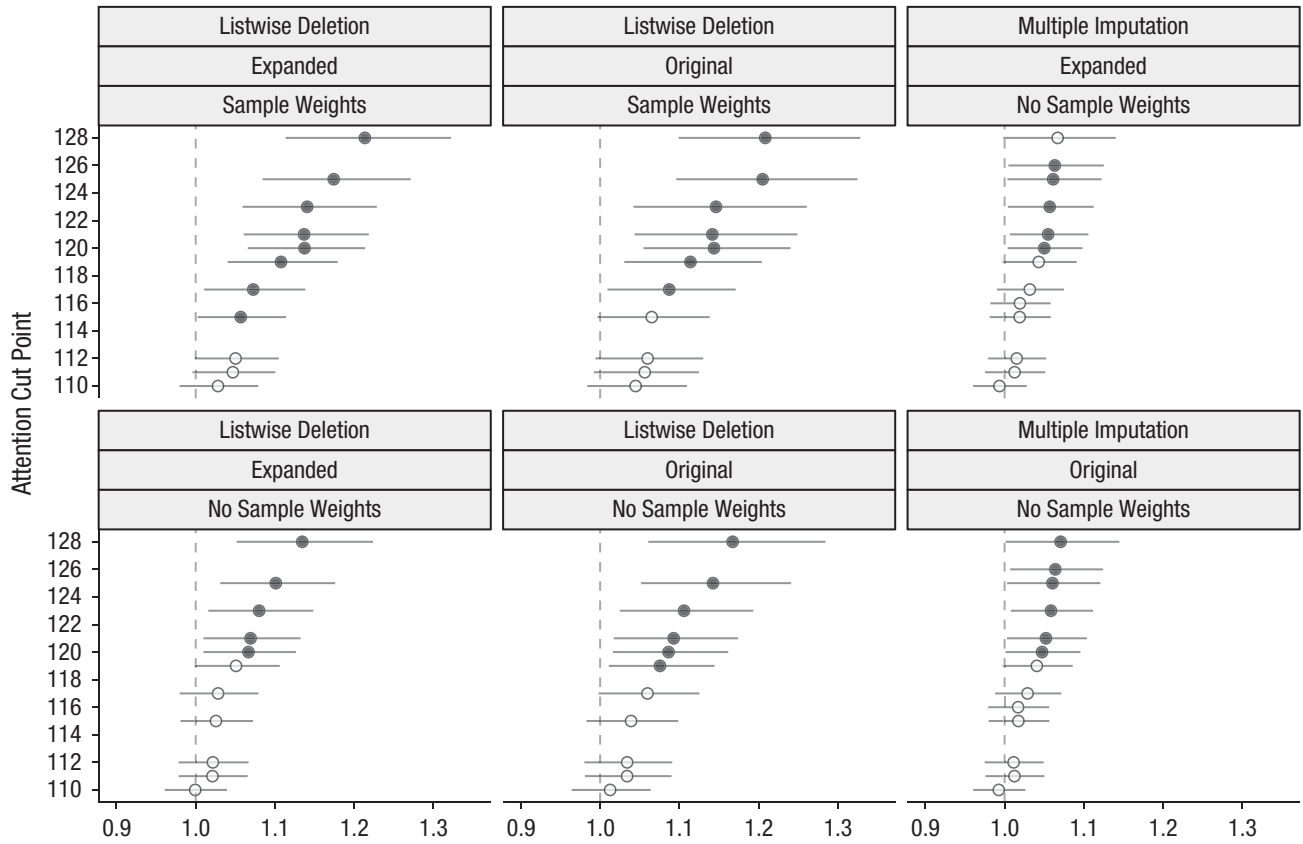
Fig. 1. Scatterplots showing the relationship between early childhood TV exposure and standardized within-sex attention score at age 7 years. The left column shows TV exposure measured at the age of approximately 1.5 years as a function of (a) standardized raw attention data at age 7 and (c) adjusted (residualized) attention score at age 7 with the effect of covariates removed. The right column shows TV measured at the age of approximately 3 years as a function of (b) standardized attention at age 7 and (d) covariate-adjusted attention at age 7. In (c) and (d), the red “x” points are adjusted on the basis of imputed covariate values. The solid blue smoothing line fits to nonmissing data only; the red dashed smoothing line fits all data (including imputed values). Point locations are slightly jittered to reduce overplotting. The shaded band around each smoothing line represents the 95% confidence interval.

Statistically significant estimates were observed in 4 of 24 (16.7%) of the models, all in the hypothesized direction. The median p value for these significant models was .027. All four of the significant estimates were produced by models using listwise deletion, and three of the four also incorporated sample weights. None of the models that used multiple imputation without sample weights yielded significance.

IPTW PSA results. Treatment-effect estimates from the propensity-score models are reported in the form of Cohen’s d . The median effect size across all IPTW models was 0.068, with 1st and 3rd quartiles of [0.005, 0.129], median $p = .253$. Overall, 100 of 384 models (26.0%) produced significant estimates in the predicted direction. Results for these models are displayed in Figure 5. Again, models that used sample weights produced significant

a

Classification Based on Within-Sex Standardized Attention
TV Exposure Measured at Age ~3



b

Classification Based on Raw Attention
TV Exposure Measured at Age ~3

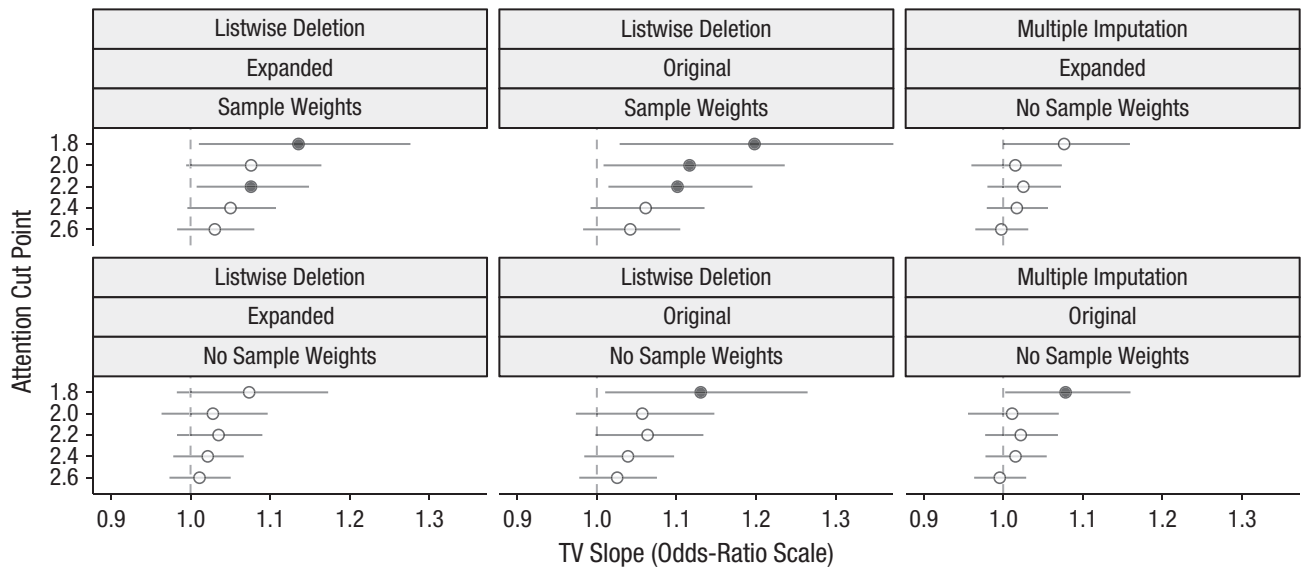
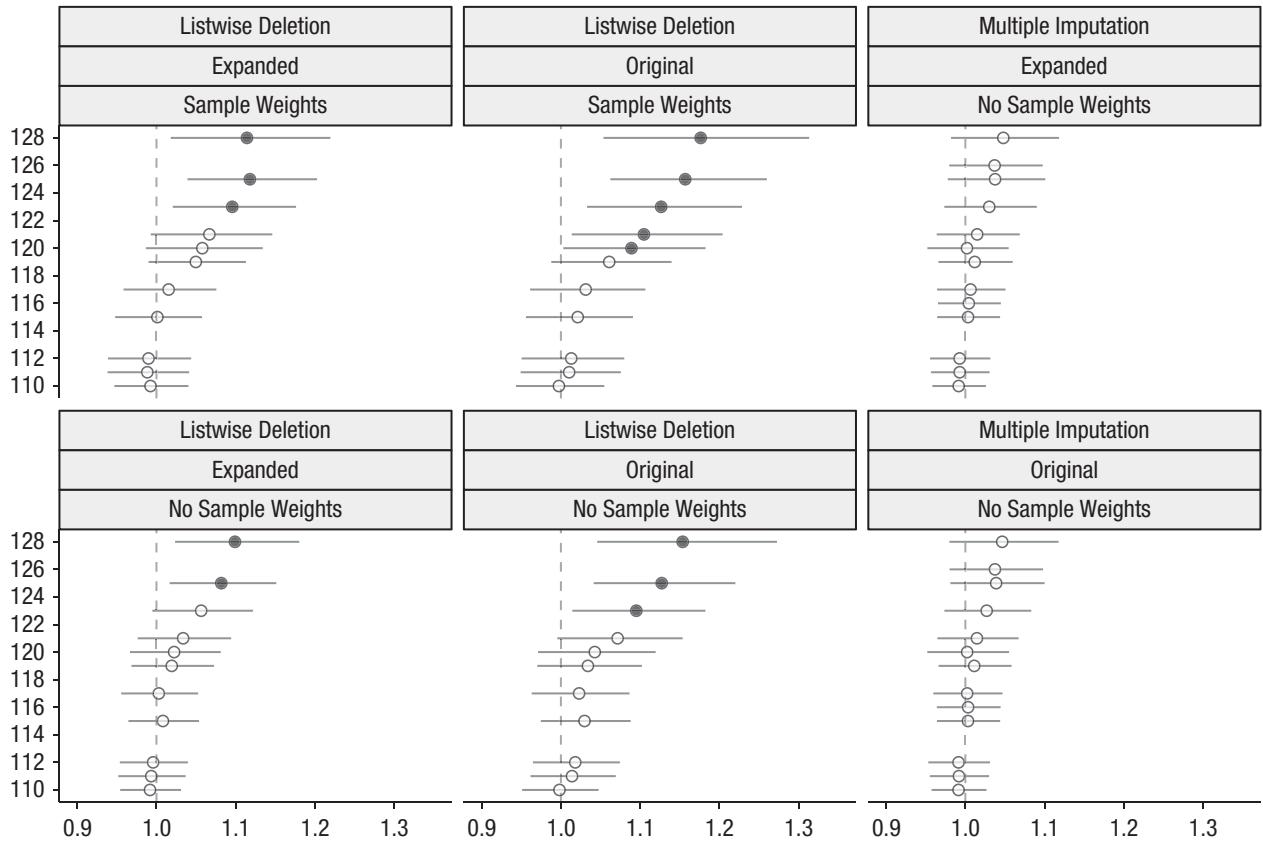


Fig. 2. (continued on next page)

C

Classification Based on Within-Sex Standardized Attention
TV Exposure Measured at Age ~1.5



d

Classification Based on Raw Attention
TV Exposure Measured at Age ~1.5

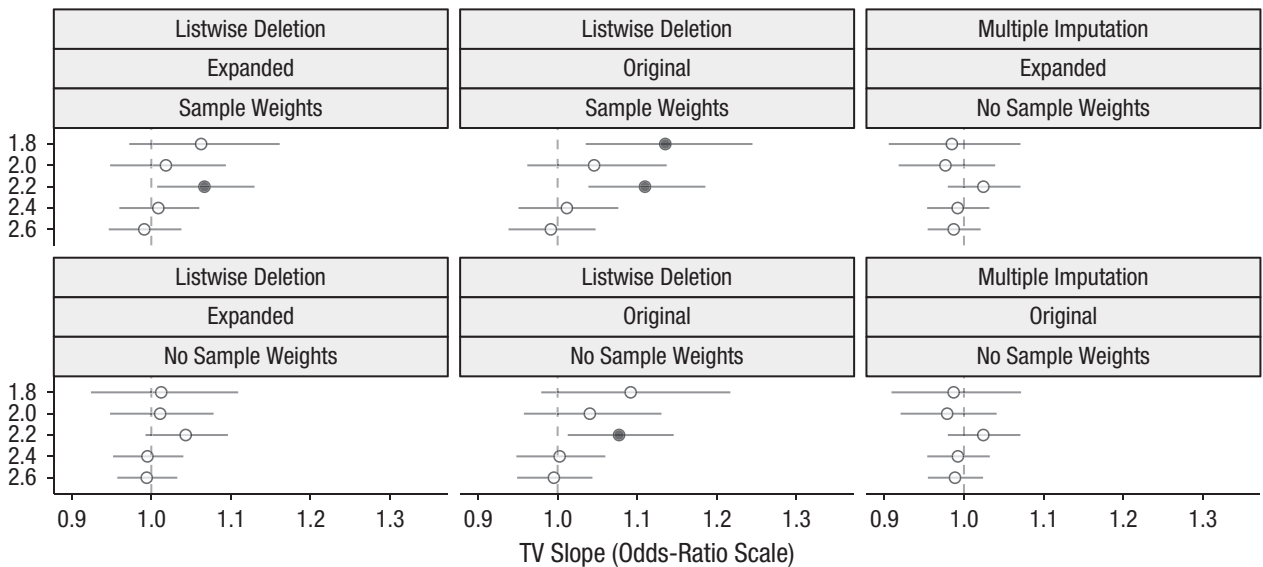


Fig. 2. Multiverse I: summary of logistic regression results. The odds ratio (OR) point estimates is shown for each attention cut point. The top row shows classifications based on within-sex standardized attention outcome and TV exposure measured at the ages of (a) approximately 3 years and (c) approximately 1.5 years, and the bottom row shows classifications based on raw attention outcome and TV exposure measured at the ages of (b) approximately 3 years and (d) approximately 1.5 years. Other model features are listed in the header to each pane. The y-axis of each panel shows the cut point defining problematic attention. The dashed vertical reference line represents no association ($OR = 1$). Error bars indicate 95% confidence intervals. Filled circles indicate significant results ($p < .05$).

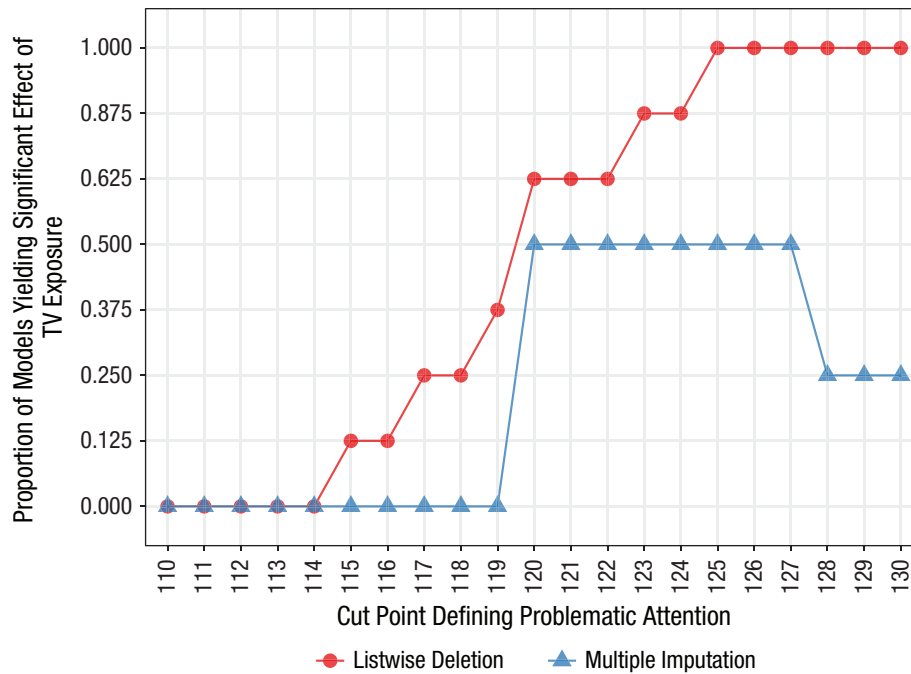


Fig. 3. Proportion of models yielding a significant effect of TV exposure as function of cut point defining problematic attention and missing-data treatment method. The figure displays results from models using the within-sex standardized attention score (rather than the raw attention score) as the basis for defining problematic attention.

results more often (80/192, or 41.7%) than those that did not (20/192, or 10.4%).

Table 4 describes how the significance of these models varied across the six cut points for high versus low TV exposure. The highest rates of significance were associated with the 50th- and 60th-percentile cutoffs.

Stratification PSA results. The median effect size for the stratification models was -0.016 , with 1st and 3rd quartiles of $[-0.041, 0.021]$, median $p = .640$. Only 1 of 240 of the stratification propensity-score models (0.4%) produced a statistically significant result, which was in the direction suggesting a beneficial effect of TV exposure. In general, the stratification models had wider standard errors and confidence intervals than the IPTW propensity-score models. Results for these models are summarized in Figure 6.

Overall summary for Multiverse II. Overall, 104 of 648 (16.0%) of models in Multiverse II produced statistically significant results in the predicted direction (indicating a harmful effect of TV exposure). One additional model was significant in the opposite direction. Looking only at models that did not use sample weights, we found that only 22 of 448 (4.9%) produced significant results. In the propensity-score IPTW models, the results varied depending on how high- and low-TV exposure was

defined, with more models showing significance when the sample was split around the 50th to 60th percentile.

Discussion

Our goal in this study was to reevaluate the claim that early exposure to TV causes attentional problems (Christakis, 2011; Christakis et al., 2004). Analyzing the same data set as Christakis et al. (NLSY79) using 848 distinct models, we found that only 166 (19.6%) produced evidence of a relationship between variables. If TV exposure truly affected attention as Christakis (2011) claimed, a substantial majority of analyses would be expected to yield significant results. Further, the superior analytic approaches should have yielded higher rates of statistical significance than the inferior ones. Instead, the opposite occurred. Looking only at the models we deemed most principled—those that did not include sample weights, discard records with missing data, or artificially dichotomize the outcome variable for logistic regression—we found that only 21 of 440 (4.8%) of the models produced significant results. Thus, our conclusion is that the claim that TV exposure harms attention is not robust and is unlikely to be true.

The most straightforward method of visualizing the relationship—the simple scatterplots presented in Figure 1—suggests a lack of compelling evidence for this

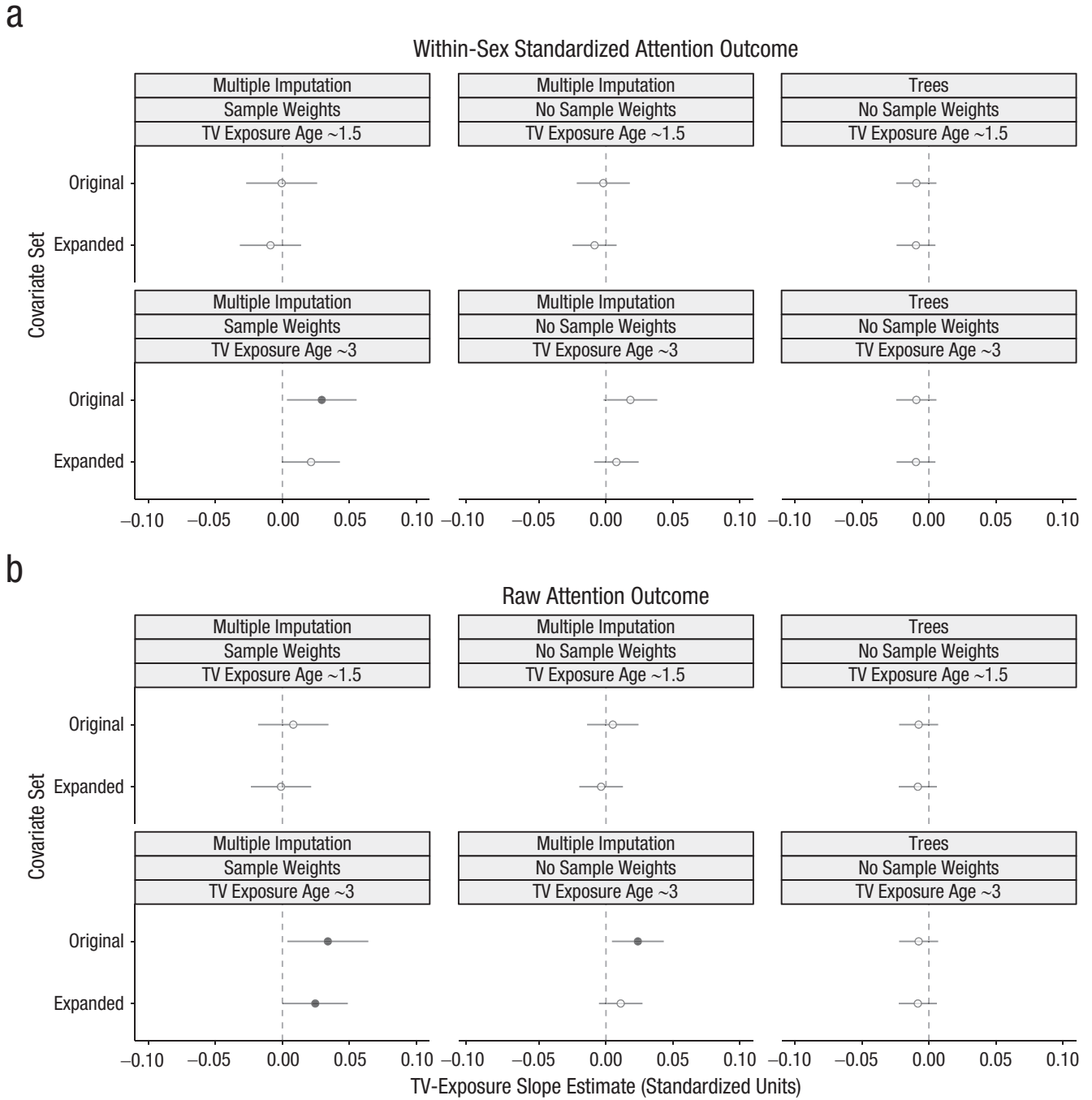


Fig. 4. Multiverse II: summary of linear regression model results. The standardized regression slope for TV exposure is shown for both the original and expanded covariate set in each model. Results are shown separately for models of the (a) within-sex standardized attention outcome and (b) raw attention outcome. Other model features are listed in the header to each pane. The outcomes are scaled so that higher scores indicate worse attention. The estimates describe the expected decrease in attention, measured in standard-deviation units, for a 1-hr increase in TV exposure. The dashed vertical reference line represents no association ($\beta = 0$). Error bars indicate 95% confidence intervals. Filled circles indicate significant results ($p < .05$).

purported relationship. Particularly when analyses included covariates, the relationship is basically a flat line—there is little visual evidence of a linear relationship in which more TV exposure leads to higher levels

of attention problems. Our formal analyses mostly underscored that point.

Under the null hypothesis, we would expect about 5% of the models to yield significance, which is almost

a

Within-Sex Standardized Attention Outcome
TV Exposure Measured at Age ~3

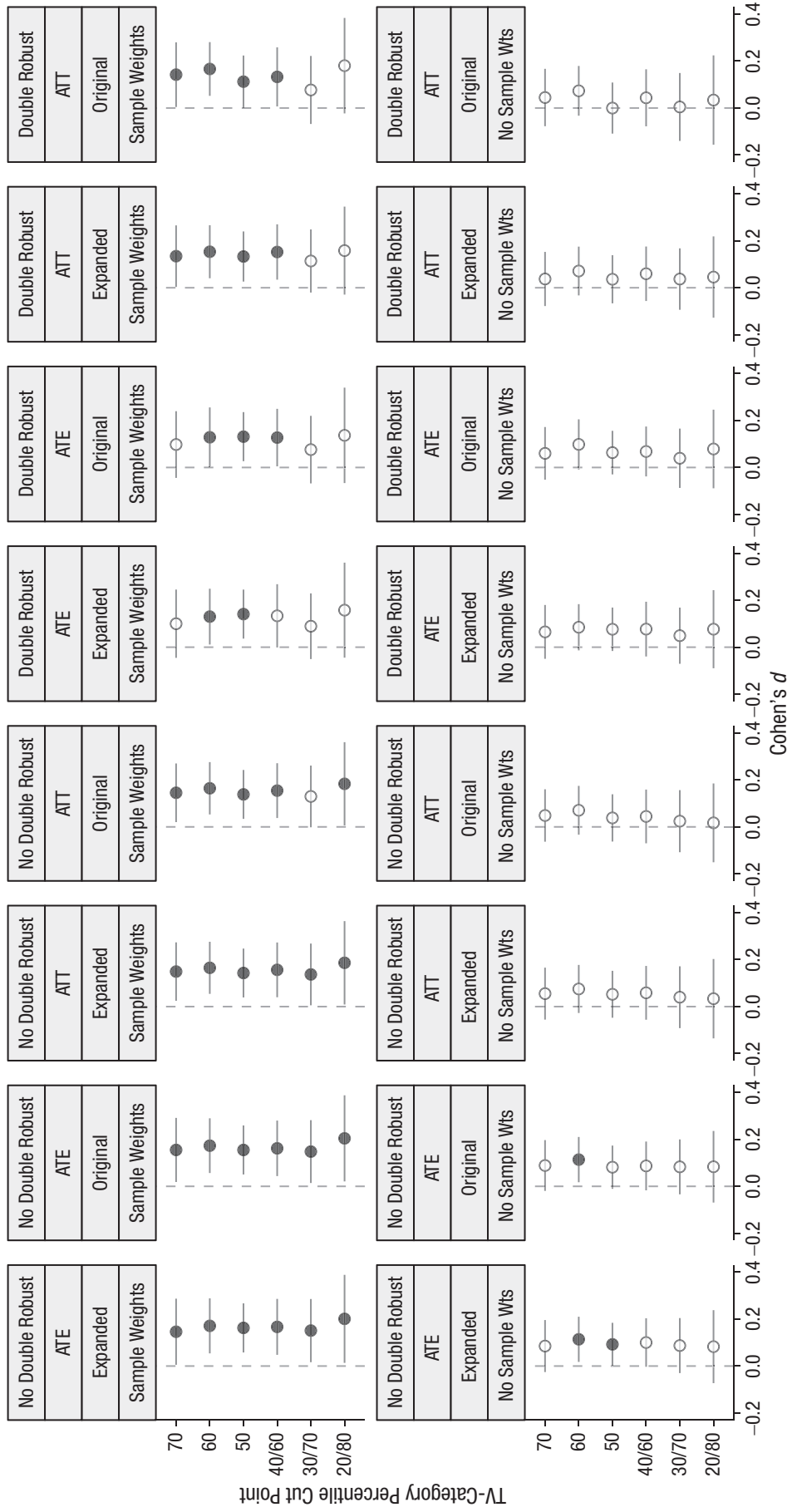


Fig. 5. (continued on next page)

b

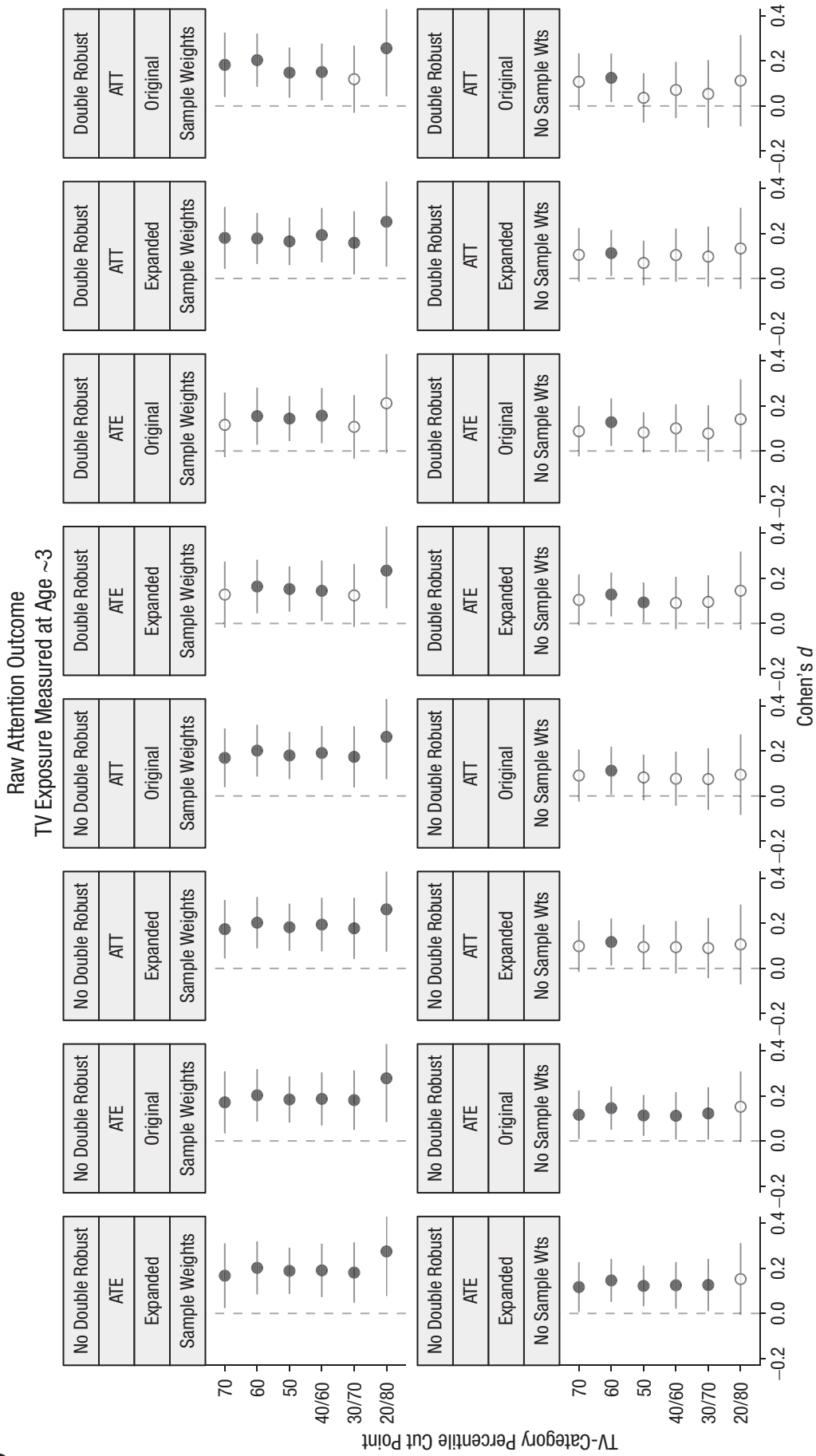


Fig. 5. (continued on next page)

C

Within-Sex Standardized Attention Outcome
TV Exposure Measured at Age ~1.5

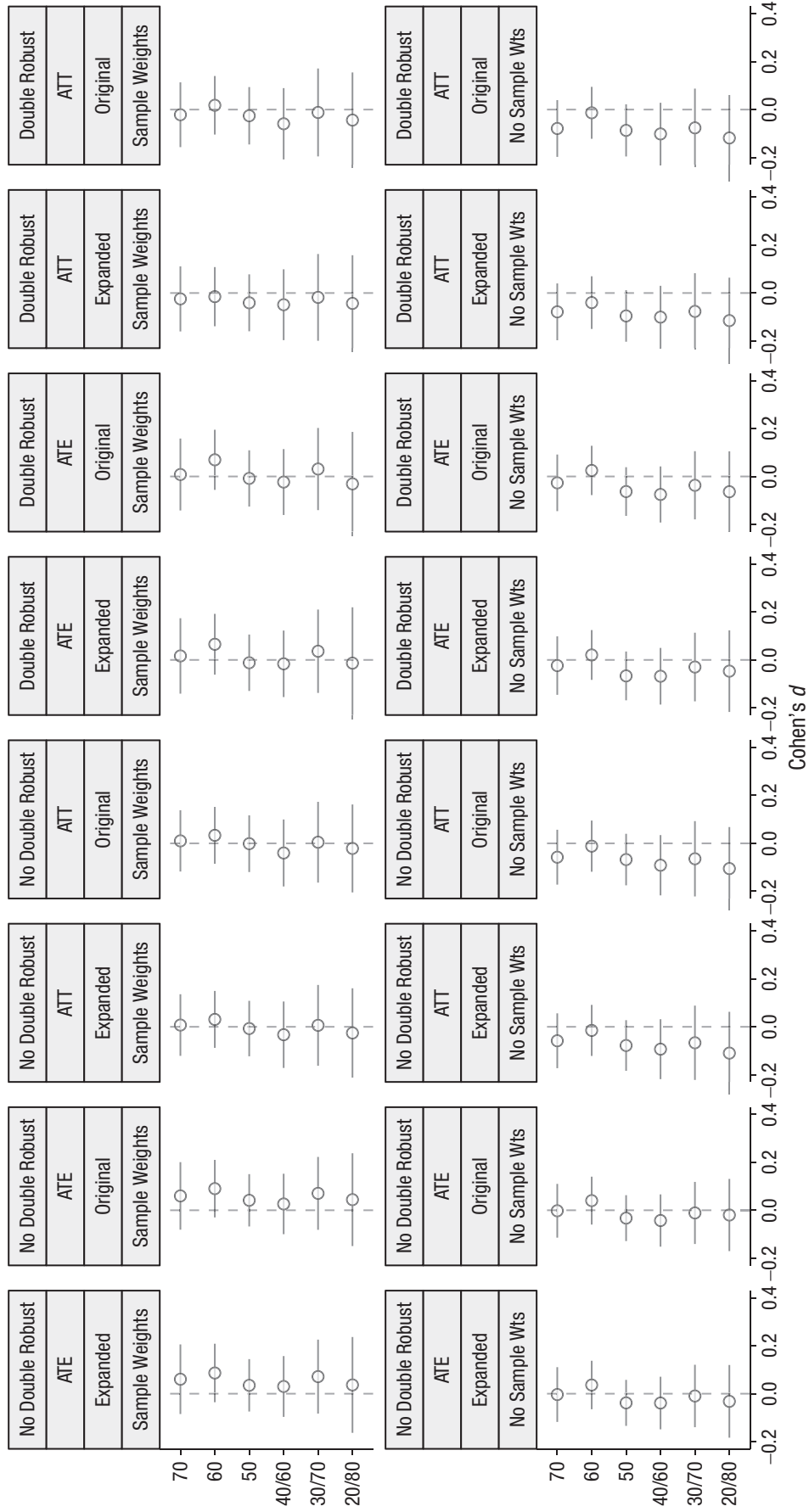


Fig. 5. (continued on next page)

d

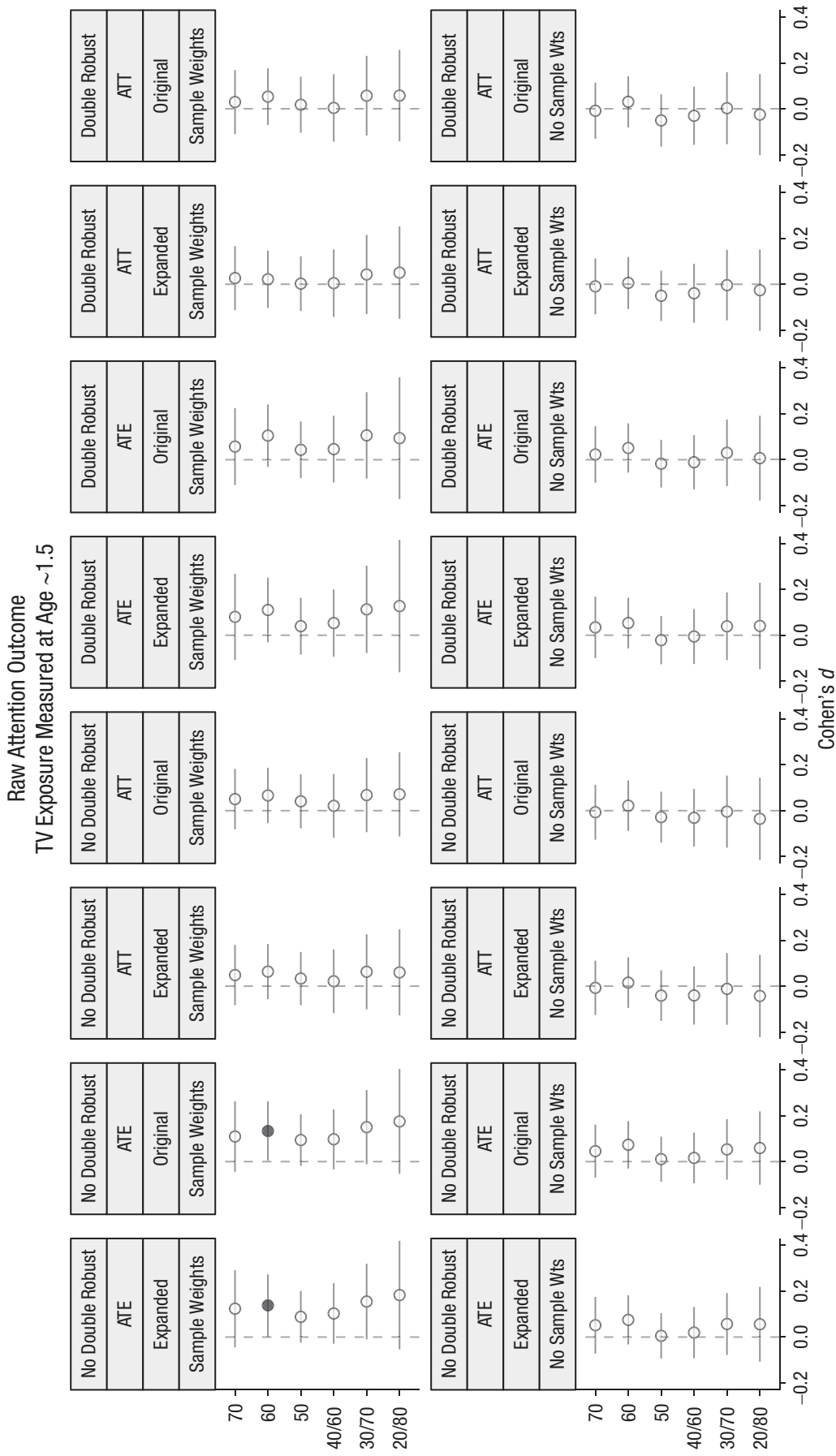


Fig. 5. Multiverse II: results of the inverse-probability-of-treatment weighting (IPTW) propensity-score analysis. Standardized effect-size (Cohen's *d*) estimates are presented for each of the percentile cut points defining the high- and low-TV groups. The top row shows models of within-sex standardized attention outcome and TV exposure measured at the ages of (a) approximately 3 years and (c) approximately 1.5 years, and the bottom rows show models of raw attention outcome and TV exposure measured at the ages of (b) approximately 3 years and (d) approximately 1.5 years. Other model features are listed in the header to each pane. The outcomes are scaled so that higher scores indicate worse attention. The dashed vertical reference line represents no association ($d = 0$). Error bars indicate 95% confidence intervals. Filled circles indicate significant results ($p < .05$). ATE = average treatment effect; ATT = average treatment effect for the treated.

Table 4. Inverse-Probability-of-Treatment Weighting (IPTW) Propensity-Score Model Results by TV-Exposure Cut Point

TV-exposure cut-point percentiles	Non-significant	Significant	Proportion significant
20/80	21	11	.344
30/70	22	10	.312
40/60	15	17	.531
50	12	20	.625
60	6	26	.812
70	18	14	.438

Note: The table includes only models measuring TV exposure at approximately age 3. *Nonsignificant* and *significant* refer to the number of models using the specified attention cut point that yielded nonsignificant versus significant results: 20/80 = below 20th versus above 80th percentile, 30/70 = below 30th versus above 70th percentile, 40/60 = below 40th versus above 60th percentile, and 50, 60, 70 = below versus above 50th, 60th, and 70th percentiles, respectively.

precisely what we found using the most principled models. Why did we find a higher percentage of significance across the broader array of analyses? We believe that the nonlinear wiggle in the scatterplots displayed in Figure 1, which is likely due to chance, triggers significance when brought into sharp relief by the analysis. That is, only certain cut points defining high versus low TV exposure (in IPTW propensity-score models) and certain cut points defining normal versus problematic attention (in logistic regression) magnify the small linear trend in a subset of the data. See the Supplementary Materials file at <https://osf.io/c895p/> for detailed post hoc analyses to support this claim.

Even if we were to cherry-pick the most alarming significant findings, the story would hardly be one worthy of concern. While the median significant *OR* from the logistic models ($OR = 1.10$) would indeed be worrisome, the magnitude of this estimate is inconsistent with the estimates from all the models that treated the outcome as continuous. The largest effect size (d) in the IPTW propensity-score models was 0.28 from a model using 20th- and 80th-percentile TV-exposure cut points, a median difference between groups of 6.3 hr of TV per day. The largest effect size from the linear regression models indicated that each hour of additional TV exposure would be associated with a 0.034 standard-deviation increase in attention problems. In real terms, this suggests that watching over 6 hr of TV per day in early childhood would not be enough to move a child from “never” to “sometimes” on even one of the five items on the hyperactivity subscale. Again, these are the largest estimates from these model families.

Our hunch at the outset of this project was that any relationship between early TV exposure and later attention

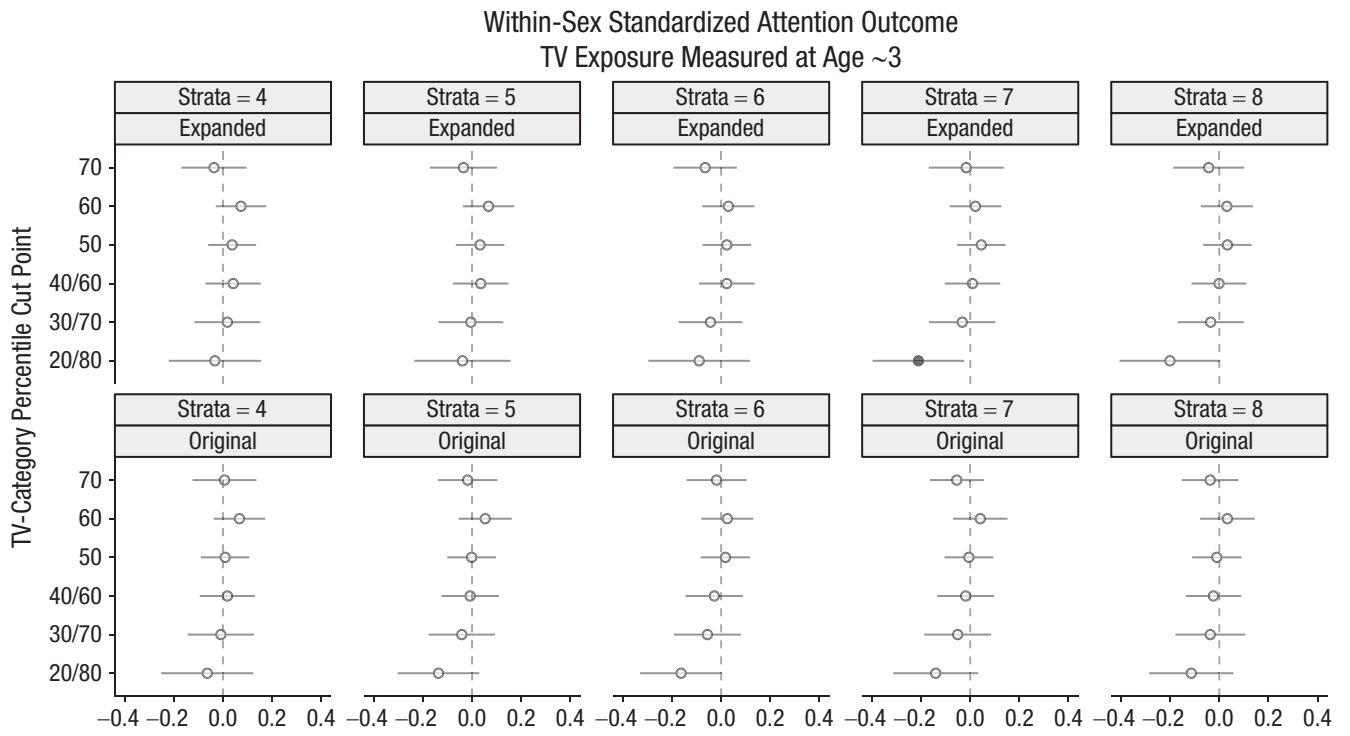
problems might be the result of the third variable of temperament. The inclusion of temperament and the other additional covariates had almost no impact on the results: 80 of 424 (18.9%) of the models using the original covariate set were statistically significant, compared with 86 of 424 (20.3%) of models using the expanded covariate set. In fact, there was little indication of a relation between TV exposure and attention to be explained at all.

One might argue that the current analysis is unnecessary because the field has already moved beyond the broad-brush claims of the original article. Recent research about screen-media use in children has become much more precise—investigating the specific effects of violent content, fantastical content, pace of scene change, and the viewer’s voluntary control of the action, among other factors (Huber et al., 2018). Notably, however, much of this research was founded on the desire to locate a mechanism for the purported negative effect of TV—an effect that our multiverse analysis calls into question. Further, although the field may have moved on to such nuanced questions, clearly the public consciousness has not; parents often continue to echo the message that TV exposure causes attention problems. We think the results of our analysis—that TV likely does not cause attention problems—bear repeating.

We also hope the current article adds to the discussion regarding the replicability crisis in inferential science. One method for increasing the reliability of research findings is the preregistration of the study design and analysis plan. Preregistration constrains researchers’ ability to endlessly reanalyze decision sets until they “discover” affirmative claims. However, preregistration does not fully address the deeper issue of model dependence, because a single analysis plan could still produce a nonrepresentative result by chance. The alternative is to make transparent the consequences of the multiple decision sets employed in an investigation. If preregistering a single analysis is good, showing the results of many possible analyses is better (with preregistration of a set of analyses arguably being best).

In summary, the multiverse analyses presented in this article used a large, nationally representative data set to ask the same question in 848 different ways: Does TV exposure in toddlerhood cause attention problems in later childhood? According to the data presented here, there is no reason to think so. We found that the TV-attention link claimed by Christakis et al. (2004) was not robust to model specification. The significance exhibited by a minority of the models (166, or 19.6%) appears to be related to overfitting a small feature of the data, and it is one that we would not expect to replicate in other samples. Overall, these data provide

a



b

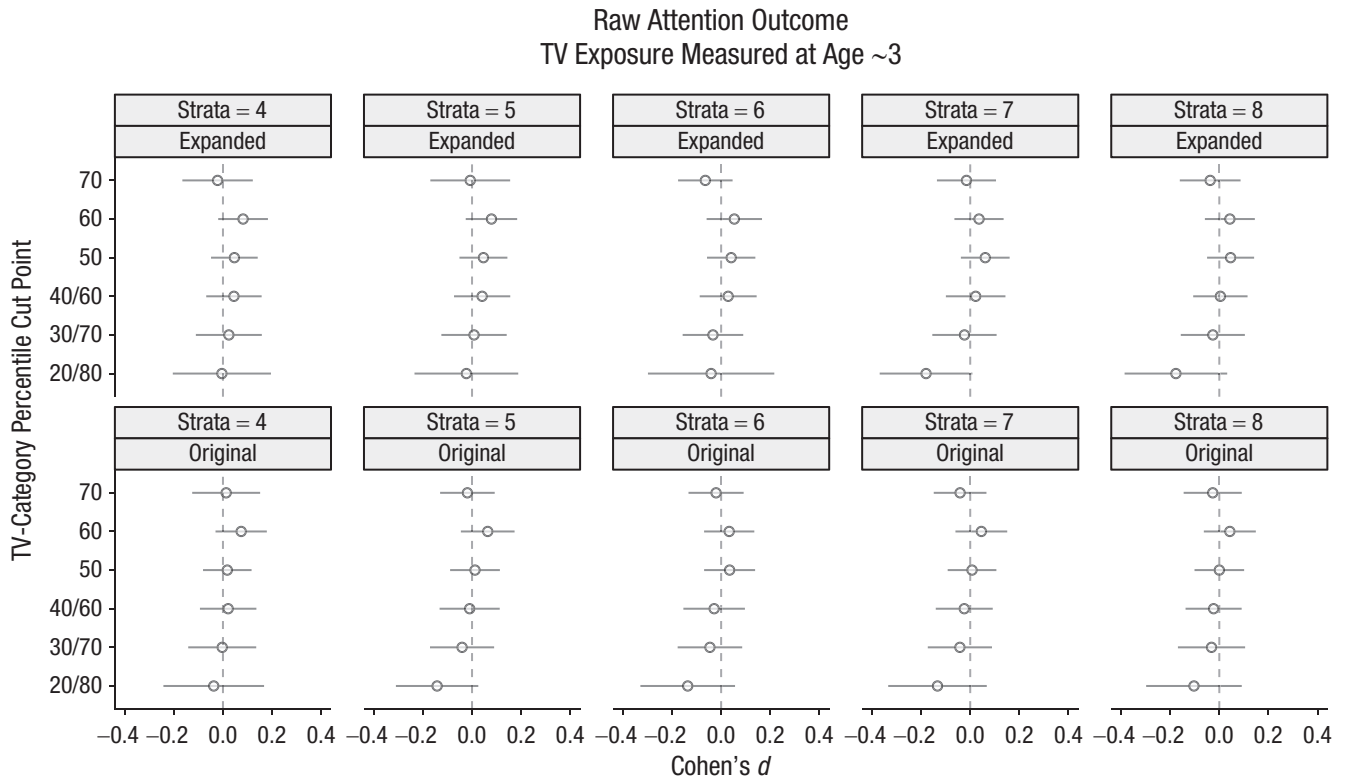
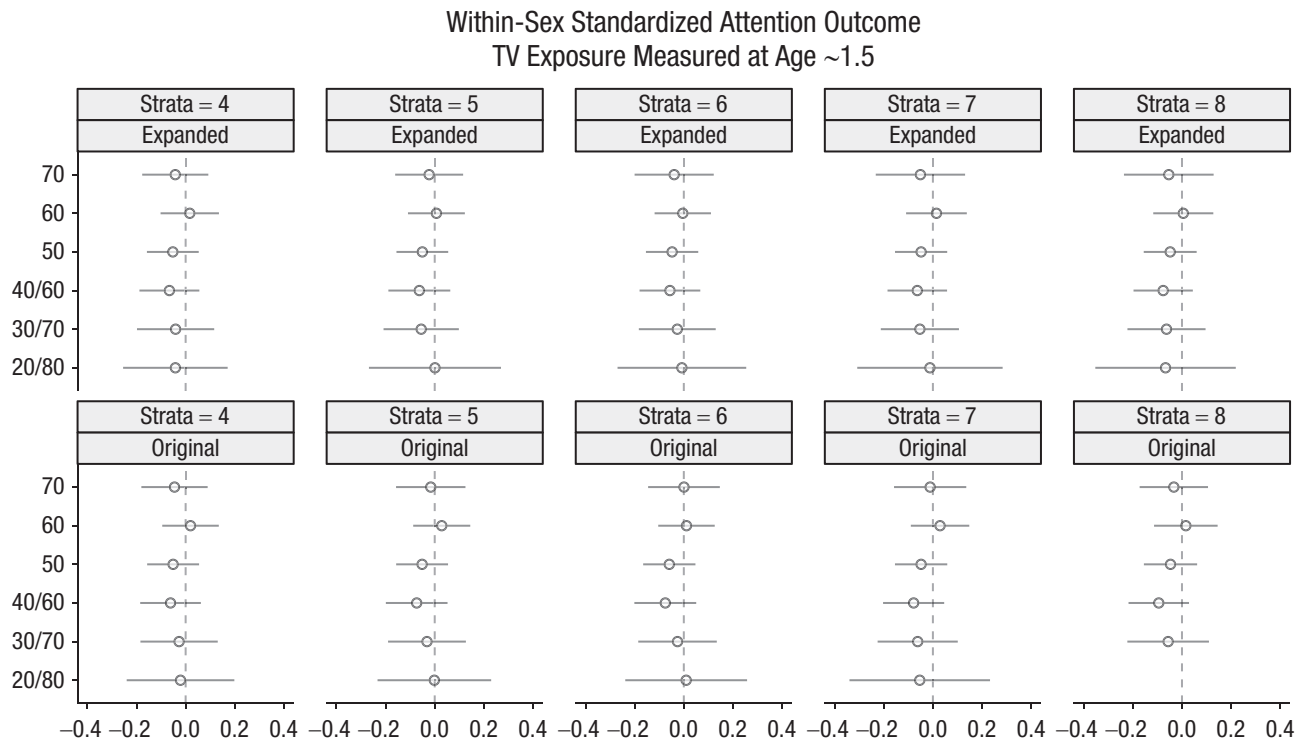


Fig. 6. (continued on next page)

C



d

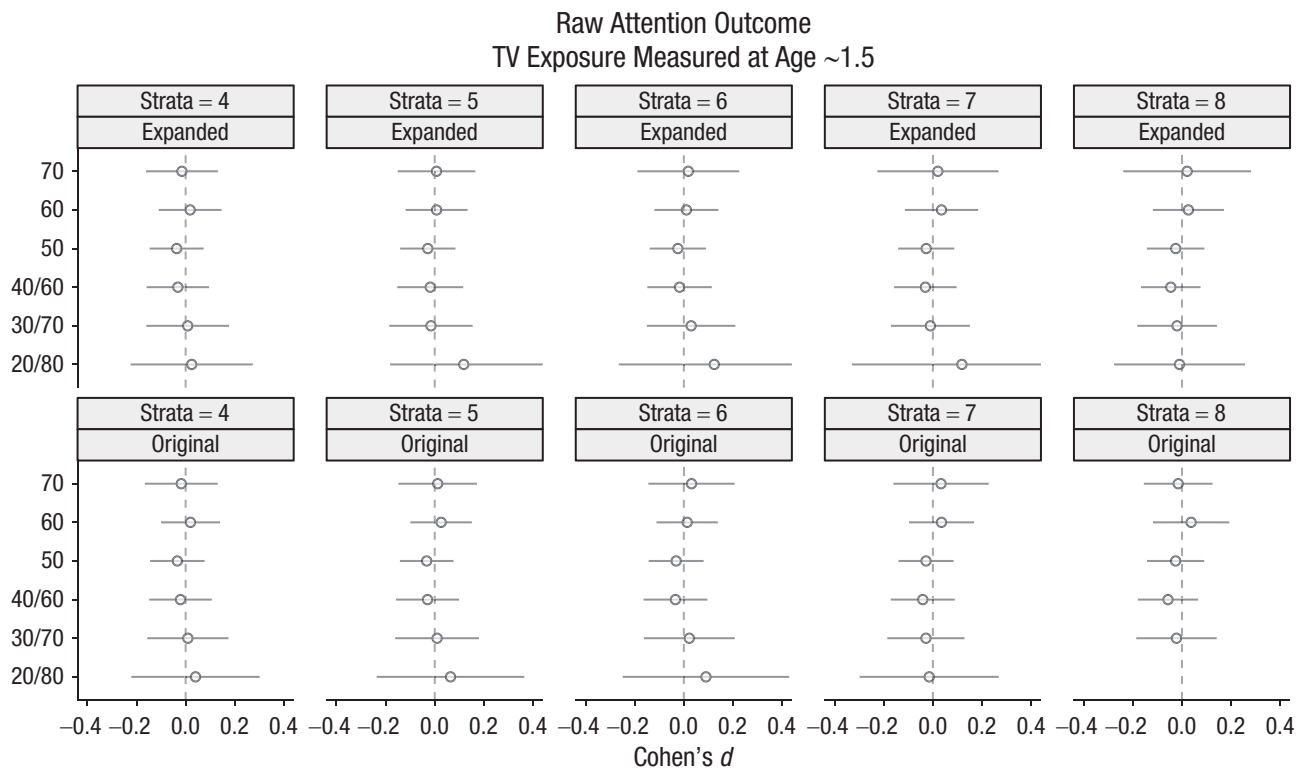


Fig. 6. Multiverse II: results of the stratification propensity-score analysis. Standardized effect-size (Cohen's d) estimates are presented for each of the percentile cut points defining the high- and low-TV groups. The top row shows models of within-sex standardized attention outcome and TV exposure measured at the ages of (a) approximately 3 years and (c) approximately 1.5 years, and the bottom row shows models of raw attention outcome and TV exposure measured at the ages of (b) approximately 3 years and (d) approximately 1.5 years. Other model features are listed in the header to each pane. The outcomes are scaled so that higher scores indicate worse attention. The dashed vertical reference line represents no association ($d = 0$). Error bars indicate 95% confidence intervals. Filled circles indicate significant results ($p < .05$).

no reason for us to believe that early TV exposure harms children's later attention. Perhaps screen media are just one more part of life that has the power to entertain, teach, confuse, distract, or inspire.

Transparency

Action Editor: D. Stephen Lindsay

Editor: D. Stephen Lindsay

Author Contributions

R. J. Brand and W. E. Dixon, Jr., jointly developed the initial idea for this study. M. T. McBee proposed the multiverse-analysis concept and was primarily responsible for writing code for data extraction and statistical analyses. M. T. McBee also created all tables and figures and organized the project's GitHub page (subsequently mirrored and registered at OSF). R. J. Brand wrote most of the first draft of the manuscript, and M. T. McBee filled in the statistical pieces. All three authors participated materially in conceptualization, development, and writing of the final draft of the manuscript. All authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Open Practices

All data and code for this study, along with a vast set of supplementary materials, have been made publicly available via OSF and can be accessed at <https://osf.io/4u69g>. The design and analysis plans for the study were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iD

Wallace E. Dixon, Jr.  <https://orcid.org/0000-0002-9260-3348>

Notes

1. Although Christakis et al. (2004) described this cut point of 120 as "1.2 standard deviations (*SDs*) above the mean" (p. 709), it is actually 1.33 standard deviations above the mean given that the standardized attention scores were constructed to have a standard deviation of 15.
2. For more information, see <https://www.nlsinfo.org/content/cohorts/nlsy97/using-and-understanding-the-data/sample-weights-design-effects/page/0/0/#practical>.
3. A prior version of this analysis also included the child's body mass index, but we removed that variable at the direction of a reviewer, who was concerned that it could be an outcome of TV exposure rather than a confounder.

References

Auerbach, J. G., Berger, A., Atzaba-Poria, N., Arbel, S., Cypin, N., Friedman, A., & Landau, R. (2008). Temperament at 7,

12, and 25 months in children at familial risk for ADHD. *Infant and Child Development*, 17(4), 321–338. <https://doi.org/10.1002/icd.579>

- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46, 399–424. <https://doi.org/10.1080/00273171.2011.568786>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Christakis, D. (2011, November). Media and children [Video]. TEDxRainier. <https://tedxseattle.com/talks/dimitri-christakis-media-and-children/>
- Christakis, D. A., Zimmerman, F. J., DiGiuseppe, D. L., & McCarty, C. A. (2004). Early television exposure and subsequent attentional problems in children. *Pediatrics*, 113(4), 708–713. <https://doi.org/10.1542/peds.113.4.708>
- Committee to Review Adverse Effects of Vaccines. (2012). *Adverse effects of vaccines: Evidence and causality*. National Academies Press.
- Foster, E. M., & Watkins, S. (2010). The value of reanalysis: TV viewing and attention problems. *Child Development*, 81(1), 368–375. <https://doi.org/10.1111/j.1467-8624.2009.01400.x>
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time* [Unpublished manuscript]. Department of Statistics, Columbia University. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Guo, S. Y., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and application* (2nd ed.). SAGE.
- Helmreich, J. E., & Pruzek, R. M. (2009). PSAgraphics: An R Package to support propensity score analysis. *Journal of Statistical Software*, 29(6). <https://doi.org/10.18637/jss.v029.i06>
- Huber, B., Yeates, M., Meyer, D., Fleckhammer, L., & Kaufman, J. (2018). The effects of screen media content on young children's executive functioning. *Journal of Experimental Child Psychology*, 170, 72–85. <https://doi.org/10.1016/j.jecp.2018.01.006>
- King, G., & Zeng, L. (2007). Detecting model dependence in statistical inference: A response. *International Studies Quarterly*, 51, 231–241. <https://doi.org/10.1111/j.1468-2478.2007.00449.x>
- Kostyrka-Allchorne, K., Cooper, N. R., & Simpson, A. (2017). The relationship between television exposure and children's cognition and behavior: A systematic review. *Developmental Review*, 44, 19–58. <https://doi.org/10.1016/j.dr.2016.12.002>
- Lotus, J. (2020, April 21). *TV linked to attention deficit*. Ovi. <https://ovimagazine.home.blog/2020/04/21/tv-linked-to-attention-deficit-by-jean-lotus/>

- Lovibond, P. F. (1998). Long-term stability of depression, anxiety, and stress syndromes. *Journal of Abnormal Psychology, 107*, 520–526. <https://doi.org/10.1037/0021-843X.107.3.520>
- Lumley, T. (2014). *Analysis of complex survey samples*. <https://www.jstatsoft.org/v09/i08/paper>
- Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760–775. <https://doi.org/10.1111/ajps.12357>
- Nabi, R. L., & Krcmar, M. (2016). It takes two: The effect of child characteristics on U.S. parents' motivations for allowing electronic media use. *Journal of Children and Media, 10*, 285–303. <https://doi.org/10.1080/17482798.2016.1162185>
- National Longitudinal Surveys. (2020, April 21). *Temperament (how my child usually acts)*. <https://www.nlsinfo.org/content/cohorts/nlsy79-children/topical-guide/assessments/temperament-how-my-child-usually-acts>
- Nature. (2016). Go forth and replicate! *Nature, 536*(7617), 373. <https://doi.org/10.1038/536373a>
- Nikkelen, S. W., Valkenburg, P. M., Huizinga, M., & Bushman, B. J. (2014). Media use and ADHD-related behaviors in children and adolescents: A meta-analysis. *Developmental Psychology, 50*(9), 2228–2241. <https://doi.org/10.1037/a0037318>
- Oliver, J. E., & Wood, T. (2014). Medical conspiracy theories and health behaviors in the United States. *JAMA Internal Medicine, 174*(5), 817–818. <https://doi.org/10.1001/jamainternmed.2014.190>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Orben, A., Dienlin, T., & Przybylski, A. K. (2019). Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences, USA, 116*(21), 10226–10228. <https://doi.org/10.1073/pnas.1902058116>
- Peck, P. (2004, April 5). Toddler TV time can cause attention problems. *WebMD*. <https://www.webmd.com/parenting/news/20040405/toddler-tv-time-can-cause-attention-problems#1>
- Posner, M. I., & Rothbart, M. K. (2018). Temperament and brain networks of attention. *Philosophical Transactions of the Royal Society B, 373*(1744), Article 20170254. <https://doi.org/10.1098/rstb.2017.0254>
- Radesky, J. S., Silverstein, M., Zuckerman, B., & Christakis, D. A. (2014). Infant self-regulation and early childhood media exposure. *Pediatrics, 133*(5), e1172–e1178. <https://doi.org/10.1542/peds.2013-2367>
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 3.6.3) [Computer software]. <https://www.R-project.org/>
- Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B., & Burgettey, L. (2017). *Toolkit for weighting and analysis of nonequivalent groups (TWANG)* (Version 1.6) [Computer software]. RAND Corp.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27–42. <https://doi.org/10.1177/2515245917745629>
- Rothbart, M. K., & Bates, J. E. (2006). Temperament. In N. Eisenberg, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology: Social, emotional, and personality development* (6th ed., Vol. 3, pp. 99–166). Wiley.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dall Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytical choices affect results. *Advanced in Methods and Practices in Psychological Science, 1*(3), 337–356. <https://doi.org/10.1177/2515245917747646>
- Smith, P. H., Dixon, W. E., Jr., Jankowski, J. J., Sanscrainte, M. M., Davidson, B. K., & Loboschefski, T. (1997). Longitudinal relationships between habituation and temperament in infancy. *Merrill-Palmer Quarterly, 43*(2), 291–304.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. <https://doi.org/10.1177/1745691616658637>
- Sullivan, E. L., Holton, K. F., Nousen, E. K., Barling, A. N., Sullivan, C. A., Propper, C. B., & Nigg, J. T. (2015). Early identification of ADHD risk via infant temperament and emotion regulation: A pilot study. *Journal of Child Psychology and Psychiatry, 56*(9), 949–957. <https://doi.org/10.1111/jcpp.12426>
- Thompson, A. L., Adair, L. S., & Bentley, M. E. (2013). Maternal characteristics and perception of temperament associated with infant TV exposure. *Pediatrics, 131*(2), e390–e397. <https://doi.org/10.1542/peds.2012-1224>
- Trzesniewski, K. H., Donnellan, M. B., & Robins, R. W. (2003). Stability of self-esteem across the life span. *Journal of Personality and Social Psychology, 84*, 205–220. <https://doi.org/10.1037/0022-3514.84.1.205>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3). <https://doi.org/10.18637/jss.v045.i03>
- Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., & Walker-Smith, J. A. (1998). Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet, 351*(9103), 637–641. [https://doi.org/10.1016/s0140-6736\(97\)11096-0](https://doi.org/10.1016/s0140-6736(97)11096-0) (Retraction published 2010, *The Lancet, 375*[9713], 445)
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE, 11*(3), Article e0152719. <https://doi.org/10.1371/journal.pone.0152719>