Running head: TV AND ATTENTION

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

Matthew T. McBee[1]

Rebecca J. Brand[2]

Wallace Dixon[1]

[1]Department of Psychology, East Tennessee State University, Johnson City, TN.

[2]Department of Psychology, Villanova University, Villanova, PA 19085

Abstract

In 2004, Christakis and colleagues published an influential paper claiming that early childhood television exposure causes later attention problems (Christakis, Zimmerman, DiGiuseppe, & McCarty, 2004), which continues to be frequently promoted by the popular media. Using the same NLSY-79 dataset (n = 2,108), we conducted two multiverse analyses to examine whether the finding reported by Christakis et al. was robust to different analytic choices. We evaluated 848 models, including logistic regression as per the original paper, plus linear regression and two forms of propensity score analysis. Only 166 models (19.6%) yielded a statistically significant relationship between early TV exposure and later attention problems, with most of these employing problematic analytic choices. We conclude that these data do not provide compelling evidence of a harmful effect of TV on attention. All material necessary to reproduce our analysis is available online via Github (https://github.com/mcbeem/TVAttention) and as a Docker container (https://hub.docker.com/repository/docker/mmcbee/rstudio_tvattention). Preprint [https://psyarxiv.com/5hd4].

Keywords: media, TV, ADHD, attention development, multiverse analysis, computational reproducibility, garden of forking paths

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

Psychological science can have a broad and deep impact on human lives. In developmental psychology 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 psychology and other fields has shown that many claims in the social and behavioral sciences literature do not hold up to re-examination (Open Science Collaboration, 2015; Camerer et al., 2018).

Once an erroneous finding has been disseminated, it seems nearly impossible to correct public misconceptions. One salient example involves Andrew Wakefield's fraudulent claim (1998, retracted) of a link between autism and the MMR vaccine (Committee to Review Adverse Effects of Vaccines, 2012; Oliver & Wood, 2014). Whether due to fraud, mismanagement, or merely chance, non-replicable findings derail scientific progress. Engaging in replication attempts and re-analyzing the robustness of reported findings are among the important strategies available for combatting this replicability crisis by determining which claims hold up to increased scrutiny and reexamination (Nature, 2016).

In this paper, we re-examine Christakis, Zimmerman, DiGiuseppe, and McCarty (2004), which claimed a positive association between television exposure in toddlerhood and attention problems at school age. Although longitudinal in nature and including a variety of control variables, the lack of randomized manipulation of TV use made it difficult to draw strong causal conclusions from these data. In our view, the provisional nature of this claim was carefully described in the paper itself. However, less qualification was used in the lead author's

subsequent public statements. For example, in a TEDx talk, the finding from this paper 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 for 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 to 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. However, if we are going to base policy and parenting guidance on the claim, we think it is important to confirm: Is it really true?

Subsequent research justifies skepticism.. A re-analysis 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 that even the direction of effect was unclear (Nikkelen, Valkenberg, Huizinga, & Bushman, 2014; see also Kostyrka-Allchorne, Cooper, & Simpson, 2017).

Given these more nuanced and updated findings, one might question whether a 16-year-old claim is worth further examination. However, it is undeniable that the meme regarding the harmfulness of screen time in general, and TV watching in particular, is still deeply embedded in popular understanding. Using Google search in April 2020 for "Does TV cause attention problems," most of the top hits, including some from reputable sites such as WebMD and whyy.org, claim a link between TV and attention problems. WebMD uses blatantly causal language in its headline ("Toddler TV Time Can Cause Attention Problems") and another site quotes Christakis as saying: "TV 'rewires' an infant's brain," and that his study shows "TV watching is a cause [of ADHD]" (Lotus, 2018). Also telling, while the original paper suggesting a link was cited 118 times in a recent two-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.

The goal of the current paper was to examine the robustness of the original claim through use of a "multiverse analysis" (Silberzahn et al. 2017; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; see also Orben, Dienlin, & Przybylski, 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. A significant finding presented in isolation can be misleading if it is not representative of larger set of possible results that could have been obtained under alternative but equally principled analytic approaches. 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 paper, we present two multiverse analyses of Christakis et al.'s (2004) original claim, using the same dataset. The first multiverse analysis employed logistic regression in conformance with the original study. The second multiverse analysis expanded the range of approaches to include linear regression and propensity score analysis techniques. See Table 1 for a conceptual map of these analyses.

--- Insert Table 1 about here ---

The first multiverse analysis closely corresponded with the original analysis, using logistic regression to predict attention problems at age 7 from TV use at age ~1.5 and ~3 years. We question the wisdom of using this technique, as it requires dichotomizing the continuous outcome variable with little justification. However, we conducted these analyses to explore whether we could replicate the original finding, and to examine the impact of modelling and data-preparation decisions within this framework. Following Christakis et al. (2004), attention problems were operationalized using the within-sex age-normed standardized hyperactivity score, which is based on a subscale of the Behavior Problems Index (BPI). These scores are nationally normed, included imputation for missing item responses, and are presented on an IQ-like metric. We also computed unstandardized ("raw") attention composite scores by taking the mean of the five relevant item responses, which were scored on a 1-3 scale. The first multiverse analysis uses both the within-sex standardized and raw scores as outcome variables, as there is no apparent reason to prefer the standardized scores in models that control for child sex and age.

In addition to *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 non-problematic and problematic levels? With no *a priori* method in the literature for determining the appropriate cutpoint, Christakis et al. (2004) 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 cutpoint from 110 to 130 on the standardized outcome and used percentile-equivalent cutpoints on the raw outcome.

The original analysis treated missing data with listwise deletion, a common but problematic technique that results in bias when the missing data mechanism is not completely random (e.g., MCAR; Little and Rubin, 2019). We explored the consequences of using listwise deletion or multiple imputation for missing data, as multiple imputation can recover unbiased estimates under a much wider set of missing data mechanisms while properly representing the uncertainty caused by the missingness. The original analysis also applied sample weights to the analysis. The NLSY documentation recommends using sample weights for estimating descriptive statistics but cautions against applying sample weights to regression or related models.[2] In our view, the analytic models should employ multiple imputation but should not use sample weights.

The second multiverse analysis employed linear regression and two forms of propensity score analysis (PSA). Unlike logistic regression, linear regression models do not require the attention variable to be dichotomized and thus eliminate the need to choose a cutpoint. This allows us to test directly the claim that for each unit change in TV use we would see a change in

[1] While Christakis et al. (2004) describes this cutpoint of 120 as "1.2 standard deviations (SDs) above the mean" (p. 709), it is actually 1.33 SDs above the mean given that the standardized attention scores were constructed to have a standard deviation of 15.

[2] https://www.nlsinfo.org/content/cohorts/nlsy97/using-and-understanding-the-data/sample-weights-design-effects/page/0/0/#practical

attention problems. Linear regression models are also a reasonable choice because they are relatively high in statistical power and efficiency. However, they require that the model be complete and correctly specified (Fox, 2016) for their estimates to be unbiased. This requires that the model include the necessary covariates, but also that the form of the mathematical relationship of each variable with the outcome is properly described by the model and that any interactions between variables are included. Specification errors involving one variable can bias the estimates for other variables. Thus, linear regression is an efficient but somewhat fragile method whose performance is based on assumptions that can be difficult to justify and harder to evaluate.

With the current dataset, we believed that propensity score analysis was the most defensible choice for estimating a causal effect. Relative to linear regression, PSA is more robust and carries a lower risk of exposure to systematic bias due to violating assumptions of the model. 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). Once the propensity scores have been estimated, they can be applied via a non-parametric technique such as matching, weighting, or stratification in order to produce virtual comparison groups that are balanced on all the covariates that were included in the propensity score model. PSA approximates random assignment of subjects to two groups (as would be seen in a true experiment; Rosenbaum & Rubin, 1983), helping to isolate TV as the potential causal variable. Further, numerous graphical and numeric diagnostic techniques are available for identifying situations that could jeopardize the results. However, it is crucial to note that propensity score

methods cannot balance the treatment and control on unobserved or unmeasured background variables as true randomization does.

Within the propensity score family of analyses, there are still many potential routes through the garden of forking paths. We explored these branching routes (see Table 1). As in Christakis et al. (2004), we included the amount of TV watched measured at two ages (~1.5 and ~3 years). We used standardized attention scores as in the original paper, as well as raw scores. Because an effect of TV watching might be different for those who watch a lot versus the average child, we calculated estimates for both the average treatment effect (ATE) and the average treatment effect for the treated (ATT).

As a technique for virtual experimentation, PSA requires dichotomizing exposure to the treatment variable into something like a treatment group and a control group. Dichotomization of continuous variables is generally not recommended because it wastes information (MacCallum, Zhang, Preacher, & Rucker, 2002). In this case, we viewed this loss of information as an acceptable tradeoff for the advantages of propensity score analysis[3]. For TV watching, one might argue that an effect would be most pronounced for the tails of the distribution (those who watch many hours a day vs. those who watch none). Choosing cutpoints at the extremes ($> 80^{th}$ percentile and $< 20^{th}$ percentile) tests this possibility, which should increase the magnitude of the difference between groups but reduces precision and power by discarding the middle 60% of the data. On the other hand, a contiguous division, such as a median split, preserves all the data but would be expected to reduce the effect size. To explore the impact of how TV watching is categorized, we ran analyses using six different cutpoints (see Table 1). If the effect of TV on

[3] Dichotomizing the response variable also wastes information, but we perceive no advantage to doing so in this case because logistic regression is no more robust, efficient, or interpretable than linear regression.

attention is linear across the range, then the effect should be proportional to the difference in median TV-watching between the groups regardless of where the cutpoint is situated.

There are also different methods for implementing PSA. We selected two techniques: inverse probability of treatment weights (IPTW) and stratification. Despite its popularity, we did not use propensity score matching because the huge multiplicity of matching algorithms and approaches would have been difficult to efficiently explore and appraise; furthermore, matching can be regarded as a subset of weighting (Imbens & Woolridge, 2009). In IPTW, the propensity scores are used to construct weights that equalize the distribution of propensity scores between the treatment and control groups – and by implication, also equalize the distribution of all of the covariates that were included in the propensity score model (Guo & Fraser, 2015). Using this method, we could estimate both the average treatment effect (ATE) and the average treatment effect for the treated (ATT). To explore the impact of using sampling weights, we ran analyses with and without them. Finally, in each model, we identified the four covariates with the largest residual imbalance statistics and gave those covariates an additional regression adjustment.[4] This is referred to as a "doubly-robust" strategy which offers protection against any remaining bias due to residual imbalance on the covariates after applying the propensity scores (Guo & Fraser, 2015). We ran analyses both with and without this doubly-robust strategy.

An alternative inferential strategy using propensity scores is to stratify (or subclassify) on them, calculate a treatment effect specific to each stratum, then combine those stratum-specific estimates to produce the average treatment effect (ATE, Guo & Fraser, 2015). Stratification allows the heterogeneity in the treatment effect to be examined across strata and provides

[4] We did not adjust for all the covariates in the doubly-robust models because doing so resulted in losing more than half the effective sample size due to listwise deletion, and there is no advantage to additional adjustment for covariates that are already well-balanced.

stratum-specific information on the success (or failure) of covariate balancing. Rosenbaum and Rubin (1983) showed that five strata are generally sufficient for removing 90% of the confounding bias. Generally, adding more strata results in better control of confounding bias at the cost of less precision and statistical power because the sample size within strata goes down as the number of strata to populate is increased. We varied the number of strata from four to eight as one of several factors to be examined.

The strength of PSA is that it approximates an experiment with random assignment; however, unlike with true random assignment, it does not balance on unobserved or omitted covariates, making the choice of covariates especially important. To explore the impact of covariate set on outcome, we conducted analyses 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).

In summary, we evaluate the claim that early childhood TV exposure is associated with increased mid-childhood attention problems. To do so, we present two multiverse analyses, using the same NLSY79 dataset, prepared in the same manner as in the 2004 paper, and employing variations of logistic regression as per the original study, and adding linear regression and propensity score models in the second multiverse analysis. In all, we examine the relation in 848 distinct ways. If the claim is true, we would expect most of the justifiable analyses to produce significant results (and in the predicted direction), and we would expect that the more rigorous, robust models (such as those treating missing data in a defensible way) would be more likely to exhibit positive results. Evaluating the outcomes across hundreds of models allows us to assess the robustness of the original claim and better understand how outcomes may be impacted by specific analytic choices. The multiverse analyses illustrate the extent to which Christakis et al.'s

claim is either representative or unrepresentative of a larger set of results that could potentially have been reported.

Method

Data

As in Christakis et al. (2004), data for the present investigation were obtained from the National Longitudinal Survey of Youth 1979 (NLSY-79), available via the NLS Investigator web interface (2020). Child data came from the NLSY79 Child and Young Adult dataset. Information on the mothers of these children came from the original NLSY79 dataset. These datasets were merged via a common ID code variable allowing mother and child data to be linked. We downloaded 340 variables from the Child and Young Adult dataset and 40 variables from the NLSY79 dataset (NLSY, 2018).

We used R version 3.6.3 (R Core Team, 2020) for data manipulation, analysis, figure generation, reporting, and automation. The *documentation* component of our project's Github page[5] presents a spreadsheet mapping our analysis variables to the variable codes and labels from the NLSY dataset. Our raw and processed analysis datasets as well as our analysis code are disclosed on this site, allowing interested readers to replicate or extend our analysis. We recommend downloading ("cloning") the repository and viewing the files locally, as html tables will not render in the browser when examining the results online. Our code and computational environment is also preserved as a Docker compute container[6], which archives the entire software toolchain (operating system + R + packages + analysis script) in a virtual machine, hedging against the possibility that future updates to any of the software components could break

[5] https://github.com/mcbeem/TVAttention
[6] https://hub.docker.com/repository/docker/mmcbee/rstudio_tvattention

the computational reproducibility of our analysis. See the Appendix for directions on reproducing our analysis in Docker.

Our variable selection process was based on the one reported in the original paper. As per Christakis et al. (2004), 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 the original study as closely as possible given the text of the original paper, 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 paper's 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 compliance with the original paper, children with severe vision or hearing impairment, as well as those with severe emotional disturbances or orthopedic disabilities were excluded. A total of 2,108 cases were extracted 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 Behavior Problems Index (BPI), which was standardized to an IQ-like metric (M = 100, SD = 15) within sex, as per the original study, which we will hereafter refer to as "within-sex standardized attention scores". However, we also retained the raw attention scores which were unadjusted for sex. The five 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 five, adding to this weekend hours per day multiplied by two, and dividing by seven. We took this measurement from three and two waves prior to the index year, such that TV was measured at approximately age 1.5 and age 3, 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 by the original article, truncating any out-of-range values of the following variables to the top of their ranges: TV use in average hours per day exceeding 16 to 16, and highest grade completed exceeding 24 to 24 (as this would imply more than eight years of post-graduate education). One high value for annual income ($839,078) was set to missing, as a comment in the NLSY codebook indicates that this value is unreliable.

The file "variable name propagation spreadsheet.xlsx" on the project Github page (under "Documentation") provides a crosswalk from our substantive, conceptual variable names to NLSY alphanumeric variable names. 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. Since these data were collected via an observational longitudinal design, confounding is a serious concern. Causal inference from observational data, in theory, is possible if the proper set of covariates are 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 those employed in the original study. They included the following: cohort (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 varied between 81 and 105 months), child's race, child's sex, the number of children of the mother living in the household, mother's highest grade completed, the cognitive stimulation and emotional support of the home (measured between ages 1 and 3), 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 by the Rosenberg Self-Esteem Scale in 1987, maternal depression as measured by the CES-D in 1992, child's gestational age at birth (centered at term), and an urbanicity indicator variable in the form of the four levels of the Statistical Metropolitan Sampling Area classification. Where applicable, all of these were extracted from the first wave of data availability to avoid conditioning on post-treatment variables, since they could have potentially biased our estimates if they were mediators or colliders (Montgomery, Nyhan, & Torres, 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 use and childhood attention. We added family income, the partner or spouse's highest level of educational attainment, an indicator variable for low birth weight (less than 2500 grams or 5 lbs 8 oz), child temperament, and an indicator that the child suffered from a health condition that limited school and play activities[7]. Rather than a continuous gestational age at birth variable, we created a binary indicator of pre-term delivery (child born before 37 weeks of gestation), as we suspected this would better capture the relevant information in this variable.

[7] A prior version of this analysis also included the child's body mass index (BMI), but we removed that variable at the direction of a reviewer, who was concerned that it could be an outcome of TV use rather than a confounder.

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 (CES-D), which was assessed only in 1992. Depending on the cohort, depression could have been assessed up to four years before birth or the same year the child was born; and self-esteem from one to five years before birth. Despite this problem of timing, we included these two variables because the original paper did. But we also expected a moderate degree of stability over time in these constructs (Lovibond, 1998; Trzesniewski, Donnellan, & Robins, 2003), which may ameliorate some concern about the timing of their measurement. We hope that including these covariates reduced 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 use, 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 television 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 dataset (NLSY Temperament, 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 she is afraid of dogs or cats, how often she cries with doctors or nurses, how often the caregiver has trouble calming the child, and how often the child cries compared to 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 in which the largest absolute standardized cross-loading was 0.133. The correlation between factors was r = -0.114. 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 use at approximately 1.5 and three years of age, and incorporated the two different sets of covariates designated above. Additional features specific to each model are described in the corresponding sections and in Table 1).

Multiverse I (Logistic Regression). First, to replicate the analysis used in the original study, we analyzed the data set using logistic regression. As already noted, Christakis et al. (2004) divided the continuous attention/behavior problems scale into typical and problematic levels of attention based on standardized attention cutpoint of 120. To determine how sensitive the original findings were to this particular cutpoint, we defined multiple dichotomous outcome variables by varying the standardized attention cutpoint from 110 to 130. For comparison

between analyses using the raw versus the standardized attention measure, we used cutpoints on the raw attention measure that were the percentile-equivalents of those on the standardized attention measure.

We fit models both with and without sample weights, using the *survey* package v. 4.0 (Lumley, 2019) to perform the weighted analysis. We also fit models both with and without multiple imputation of missing data, using the *mice* package, v. 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 cutpoints) x 2 (outcomes) x 2 (TV ages) x 2 (covariate sets) x 2 (sample weights)] and multiple imputation yielded 168 models [21 (attention cutpoints) x 2 (outcomes) x 2 (TV ages) x 2 (covariate sets)], for a combined total of 504 logistic regression models. However, because of sparseness on the attention outcome (particularly the raw version), frequently the imposition of two adjacent cutoffs (e.g, 121 and 122) would 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 Propensity Score Analysis

***Linear Regression*.** These models estimated the linear relationship between TV use, measured at approximate ages 1.5 and 3, and the mid-childhood 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 fit 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 fit 16 models [2 (outcomes) x 2 (TV ages) x 2

(covariate sets) x 2 (sample weights)], whereas using multiple imputation we fit 8 models [2 (outcomes) x 2 (TV ages) x 2 (covariate sets)], for a total of 24 linear regression models.

Propensity Score Analyses. Finally, we conducted propensity score analyses using two techniques of incorporating the propensity scores (IPTW vs. stratification). We ran analyses using both the raw and within-sex standardized versions of the outcome at age ~ 1.5 years and age ~3. To explore the impact of how hours of TV are dichotomized into "high" and "low" groups we ran analyses using six different percentile cutpoints to define the high and low TV groups as follows:

- Below 20th percentile / Above 80th percentile (20/80)
- Below 30th percentile / Above 70th percentile (30/70)
- Below 40th percentile / Above 60th percentile (40/60)
- Below 50th percentile / Above 50th percentile (50)
- Below 60th percentile / Above 60th percentile (60)
- Below 70th percentile / Above 70th percentile (70)

Where possible, we ran analyses with and without a doubly-robust strategy, with and without sampling weights, and for both the ATT and the ATE. In all the propensity score analyses, we used boosted classification trees (as implemented in the *twang* package, v. 1.6, Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017) to estimate the propensity scores, using bagging and cross-validation to prevent overfitting. Missing data on covariates is 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 cutpoints) x 2 (outcomes: raw vs.

standardized) x 2 (TV ages: 1.5 vs. 3) x 2 (covariate sets) x 2 (treatment effects: ATT vs. ATE) x 2 (sample weights) x 2 (doubly-robust vs. not)]. The *survey* package (Lumley, 2019) was used to estimate the treatment effect after applying IPTW weights.

***Stratification*.** Two hundred and forty stratification propensity score models were computed [five different numbers of strata (4, 5, 6, 7, or 8), which were fully crossed with 6 (TV cutpoints) x 2 (outcomes) x 2 (TV ages) x 2 (covariate sets)]. Neither sample weights nor the doubly-robust approach could be implemented in the stratification models, nor could these models estimate the average treatment effect for the treated (ATT). We used the *PSAgraphics* package (v 2.1.1; Helmreich & Pruzek, 2009) to perform the stratified analysis, and calculated *p*-values for the treatment effect estimates using the normal approximation.

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

Results

Obviously, space limitations prevent us from displaying detailed results from all models in this paper. However, the Github repository "Results" directory contains a subfolder for every analysis conducted, which includes descriptive statistics, diagnostic tables and plots, and formatted model results, which were produced using the *stargazer* package (v. 5.2.2; Hlavac, 2015).

Descriptive statistics

Tables 2 and 3 provide descriptive statistics for the continuous and categorical variables, respectively. The Github site contains descriptive statistics specific to each of the analysis models (for example, by low- and high-TV groups created to enable certain types of analysis). The scatterplots presented in Figure 1 illustrate the relationship between early TV consumption

(at both ~1.5 and ~3 years) and later attention measured at age ~7 years (standardized within sex). The top row represents the relation without covariates and the bottom row after removing the influence of covariates. Because missing data on the covariates dramatically reduced the sample size for the available analyses, the bottom row displays imputed data taken from the first (of 10) multiple imputations with red "x" symbols. Figure 1 also contains non-parametric smoothed regression lines to help visualize the relationship between TV and attention. The solid blue line fits to complete (non-imputed) data only, while 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 and attention. The only relation evident from the smoothed trajectory is a slight non-linear "wiggle" in the 2-6 hours per day range of TV use. This non-linearity 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 use and attention could represent a combination of confounding and missing data bias.

--- Insert Tables 2 & 3 and Figure 1 about here ---

Multiverse I Results (logistic models). For each analysis, we report the odds ratio (OR) of the relationship between TV consumption at ages ~1.5 and ~3 and the probability of being in the "problematic" category of attention, after controlling for covariates. Odds ratios greater than one indicate a higher risk of being classified into the "problematic attention" category. We vary the threshold for "problematic" attention from 110 to 130. Results are summarized in Figure 2. Effect size point estimates and confidence bounds are given in odds ratio (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 0.213. Overall, 61/200 models (30.5%) produced significant estimates in the predicted direction, and none in the opposite direction.

--- Insert Figure 2 about here ---

As shown in the figure, the results of the logistic regression analysis are highly sensitive to the choice of attention cutpoint and other features of the analysis. Statistically significant estimates of the relation between TV measured at age ~3 and the probability of attention problems begin to be observed for standardized attention cutpoints of 115 or above, and for TV measured at age ~1.5 at 120 and above.

Statistical significance occurred at a higher rate when sample weights were applied (31/64, 48.4%) than when they were not (30/136, 22.1%). When considering only those models with no sample weights[8], a much higher number yielded significance under listwise deletion (18/64, 28.1%) than under multiple imputation (12/72, 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 one. The median *p* value for these twelve significant tests was $p = 0.034$, and their median OR was 1.060. Further, we note that we were not able to exactly replicate the values reported by Chrisakis at al. (2004) 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 was measured at age ~3; Christakis et al. reported 1.09 [1.03, 1.15] for this condition. We estimated OR =

[8] There were no models with multiple imputation and sample weights, as these could not be combined.

1.058, 95% CI [0.987, 1.134], $p = 0.114$ when TV was measured at age ~1.5; whereas the original paper reported OR = 1.09 [1.02, 1.16]. We cannot explain these discrepancies.

Multiverse II results

Linear regression models. We report standardized (with respect to *y*) regression coefficients for the effect of TV on attention, such that the estimates represent the expected standard deviation change in attention given a one-hour change in TV use. 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 3A. The median estimate for these models was $b = -0.002$ with 1st and 3rd quartiles of [-0.008, 0.013], median $p = 0.335$.

--- Insert Figure 3A about here ---

Statistically significant estimates were observed in 4/24 (16.7%) of the models, all in the hypothesized direction. The median *p* value for these significant models was $p = .027$. 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 but did not incorporate sample weights yielded significance; further, all their estimates were in the 'wrong' direction (e.g, TV helps attention).

IPTW propensity score analysis results. We report Cohen's *d* effect sizes for the treatment effect estimates from the propensity score models. Note that the effect sizes for models using different TV cutpoints to define the virtual 'treatment' and 'control' groups are not equivalent, as they are based on varying degrees of difference in TV use between groups. The median effect size across all IPTW models was $d = 0.068$, with 1st and 3rd quartiles [0.005, 0.129] and median $p = 0.253$. Overall, 100 / 384 models (26.0%) produced significant estimates,

and the direction of all of these was that TV has a harmful effect on attention. Results for these models are displayed in Figure 3B.

--- Insert Figure 3B about here ---

Only 20/192 (10.4%) of these models that did not include sample weights produced significant results, compared with 80/192 (41.7%) of the models that included sample weights. Table 3 describes how the significance of these models varied across these cutpoints. The highest rates of significance were associated with the 50th and 60th percentile cutoffs.

--- Insert Table 3 about here ---

***Stratification propensity score analysis results*.** The median Cohen's *d* effect size for the stratification models was d = -0.016 with 1st and 3rd quartiles [-0.041, 0.021] and median p = 0.640. Only 1/240 of the stratification propensity score models (0.4%) produced a statistically significant result, and its point estimate indicated 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 3C.

--- Insert Figure 3C about here ---

***Overall summary for multiverse II*.** Overall, 105/648 (16.2%) of models in multiverse II produced statistically significant results, and only 22/448 (4.9%) of those not incorporating sample weights produced significant results. All significant results except one indicated a harmful effect of TV on attention. In the propensity score IPTW models, the results varied over the TV cutpoint used to define the 'treatment' and 'control' groups. Only 2/324 (0.6%) of the models measuring TV exposure at age ~1.5 produced significant results, whereas 103/324 (31.8%) of the models measuring TV at age ~3 produced significance.

Discussion

The broad goal of this paper was to re-evaluate the claim that early TV watching causes attentional problems (Christakis et al., 2004; Christakis, 2011). Over both multiverse analyses, only 166/848 (19.6%) of the models produced evidence of a relationship between variables. Further, only 21/440 (4.8%) of the models that we deemed most principled – those not including sample weights, not discarding records with missing data, and not artificially dichotomizing the outcome variable for logistic regression – produced significant results. If this dataset contained evidence of a causal association between TV and attention, the superior analytic approaches should have yielded higher rates of statistical significance. Instead, the opposite occurred, and we note that this significance rate is nearly precisely what one would expect under the null hypothesis given a 5% alpha criterion. Thus our conclusion is that the claim 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 purported relationship. Particularly when including covariates, the relationship is basically a flat line – there is little visual evidence of a linear relationship in which more TV leads to higher levels of attention problems. Our formal analyses mostly underscore that point.

Why Were Some Models Significant? We performed some 'post-mortem' analyses to better understand why some of these models detected a relationship between variables. Our explanation is that the nonlinear "wiggle" in the scatterplots displayed in Figure 1 can trigger significance if it is brought into sharp relief by the analysis. Our argument rests on a few observations. As Figure 1 indicates (comparing left panels to right panels), this feature is visually more pronounced when TV is measured at age ~3 than at age ~1.5. Over both multiverse analyses, only 19/424 (4.5%) of the age ~1.5 models were significant. compared to 147/424

(34.7%) of the age ~3 models. Furthermore, panels C and D of the figure illustrates how this "wiggle" is visually diminished by imputation of missing data; an unconfounded comparison reveals that 19/72 (26.4%) of models using listwise deletion were significant compared to 12/80 (15.0%) of those using multiple imputation.

The pattern of results in the IPTW propensity score models revealed that the likelihood of significance varied across TV cutpoints, as shown by Table 3. The highest significance rates occurred at 50th and 60th percentile cutpoints (2.86 and 3.43 hours of TV per day, respectively) for the models in which TV was measured at age ~3. This is consistent with our hypothesis. Figure 4 displays a magnified view of the nonlinear 'wiggle' and indicates how the various TV percentile cutpoints for these models aligned with this feature. The 50th and 60th percentile cutpoints, which had the highest significance rates, placed the dividing line between low and high TV groups almost precisely in the center of this nonlinear feature of the data, and resulted in the largest precision-weighted difference in the means between these groups.

--- Insert Figure 4 about here ---

The results of the logistic regression models in multiverse I also support our hypothesis. As shown in Figure 2 and Table 4, the significance of these models was strongly related to the attention cutpoint defining the 'normal' and 'problematic' attention groups. We believe that higher cutpoints in these models magnify the nonlinear feature of the data to make it more consistent with a TV-attention relationship. Figure 5 plots the proportion of cases in the 'problematic' attention category by TV use at age 3 (which has been categorized into bins to permit rate calculations) for twelve different attention cutpoints. The nonlinearity can be easily observed in the pattern of dots in the upper left panel. Each panel of the figure displays a fitted linear regression line and shaded confidence interval, which roughly represents the performance

of a logistic regression model using that cutpoint. The *p*-value for the slope coefficient of those regression lines is displayed in each panel. At low cutoffs for defining problematic attention, the nonlinear configuration of points reduces the slope of the fitted line and, more importantly, adds uncertainty regarding the slope, increasing the *p*-value for the test. As the cutoff defining problematic attention increases, the base rate of attention 'problems' is reduced, and the points migrate downward. This alters the pattern of points, making them more consistent with a linear trend. As the cutpoint surpasses 120, the *p*-value for the slope becomes significant and remains so through the highest cutpoint examined. This pattern is consistent with the results of the logistic models.

--- Insert Figure 5 about here ---

Even the Worst Case is Not So Bad. Our strong suspicion is that the nonlinear feature in the scatterplot that we have identified as the culprit for some of the significant models is likely a chance feature of the data that would be unlikely to replicate in future samples. Furthermore, 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 odds ratio 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 in the IPTW propensity score models was $d = 0.28$ from a model using 20th and 80th percentile TV cutpoints, a median difference between groups of 6.3 hours of TV per day. An effect size this small would be difficult to notice, as two distributions differing by this amount exhibit a 88.9% overlap[9]. In real terms, this suggests that watching over six hours of TV a day in early childhood would not be enough to move a child from a "never" to a "sometimes" on even one of the five

[9] See https://rpsychologist.com/d3/cohend/ for an interactive visualization of Cohen's *d* effect sizes

items on the hyperactivity subscale. This estimate is roughly consistent with the largest effect size from the linear regression models, which indicated that each hour of additional TV watching would be associated with a 0.034 standard deviation increase in the attention outcome. By this estimate, it would take an increase of 5.9 hours per day of TV watching to achieve a "small" d = 0.2 effect on attention. 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-watching 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. Whereas 80/424 (18.9%) of the models using the original covariate set were statistically significant, compared with 86/424 (20.3%) of models using the expanded covariate set. In fact, there was little indication of a relation between TV 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 from the original paper. Recent research about screen media use in children has gotten 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 onto such nuanced questions, clearly the public consciousness has not, with parents often continuing to echo the message that TV 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 paper 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 iterate over decision sets until "discovering" affirmative claims. However, preregistration does not fully address the deeper issue of model dependence, because a single analysis plan could still produce a non-representative result due to 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 pre-registration of a *set* of analyses arguably being "best").

In summary, the multiverse analyses presented in this paper used a large, nationally representative dataset to ask the same question in 848 different ways: Does TV watching 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 (166, 19.6%) of the models appears to be related to overfitting a small feature of the data, and one that we would not expect to replicate in other samples. Overall, these data provide no reason to believe that early TV harms later attention. Perhaps screen media is just one more part of life that has the power to entertain, teach, confuse, distract, or inspire.

References

Auerbach, J. G., Berger, A., Atzaba-Poria, N., Arbelle, 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*. https://doi.org/10.1038/s41562-018-0399-z

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

Christakis, D. A. (2011, November). *Media and children* [Video]. TEDxRainier. https://tedxseattle.com/talks/dimitri-christakis-media-and-children/

Committee to Review Adverse Effects of Vaccines. (2012). *Adverse effects of vaccines: Evidence and causality*. Washington, DC: 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

Fox, J. A. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). SAGE: Thousand Oakes, CA.

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*. Retrieved from 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.). Thousand Oaks, CA: SAGE.

Helmreich, J. E. & Pruzek, R. M. (2009). PSAgraphics: An R Package to support propensity score analysis. *Journal of Statistical Software, 29*(6), 1-23. https://doi.org/10.18637/jss.v029.i06

Hlavac, M. (2015). *Stargazer: Well-formatted regression and summary statistics tables.* [R package]. Version 5.2.2. http://CRAN.R-project.org/package=stargazer

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

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1). 5-86. https://doi.org/10.1257/jel.47.1.5

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 31ehavior: A systematic

review. *Developmental Review*, *44*, 19-58. https://doi.org/10.1016/j.dr.2016.12.002

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

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. (2019). *Survey: Analysis of complex survey samples* [R package]. Version 4.0.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19-40. https://doi.org/10.1037//1082-989X.7.1.19

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 US parents' motivations for allowing electronic media use. *Journal of Children and Media*, *10*, 285-303. https://doi.org/10.1080/17482798.2016.1162185

NLS Investigator (2020, April 19). Retrieved from https://www.nlsinfo.org/investigator/pages/search.jsp?s=NLSY79

National Longitudinal Survey of Youth – Temperament (How My Child Usually Acts) (2020, April 21). Retrieved from https://www.nlsinfo.org/content/cohorts/nlsy79-children/topical-guide/assessments/temperament-how-my-child-usually-acts

Nature Editorial Staff. (2016). Go forth and replicate! *Nature*, *536*, 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. http://dx.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), aac4716–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, 116*(21), 10226–10228. https://doi.org/10.1073/pnas.1902058116

Posner, M. I., & Rothbart, M. K. (2018). Temperament and brain networks of attention. *Philosophical Transactions of the Royal Society B, 373*, 20170254. https://doi.org/10.1098/rstb.2017.0254

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

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

Ridgeway, G., McCaffrey, D., Morral, A. Griffin, B., & Burgettey, L. (2017). *Toolkit for weighting and analysis of nonequivalent groups (TWANG)* [R package]. Santa Monica, CA: RAND Corporation.

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

Rosenbaum, P.R., & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55. https://doi.org/10.1093/biomet/70.1.41

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.), 3,* 99-166. John Wiley & Sons Inc.

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 dataset: 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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22,* 1359-1366. https://doi.org/10.1177/0956797611417632

Smith, P. H., Dixon Jr, W. E., Jankowski, J. J., Sanscrainte, M. M., Davidson, B. K., & Loboschefski, T. (1997). Longitudinal relationships between habituation and temperament in infancy. *Merrill-Palmer Quarterly*, *46*, 291-304.

Steegen, 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. doi 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), 1-67. https://doi.org/10.18637/jss.v045.i03

Wakefield, A.J., Murch, S.H., Anthony, A., Linnell, J., Casson, D.M., Malik, M., ... & Walker-Smith, J.A. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet*.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. PLOS ONE, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Table 1
Conceptual map of analytic decisions for the multiverse analyses

Feature	Levels
Multiverse I	
Logistic regression (504 total, 200 non-redundant)	
Attention cutpoint	110, 112, 114, …., 130 *(21 levels)*
Outcome variable	Within-sex standardized raw attention
Treatment of missing data	Listwise deletion vs multiple imputation
Sample weights	Incorporated vs not incorporated
Covariate set	Original vs expanded
TV age	~1.5 vs ~3
Multiverse II	
Linear regression (24 total)	
Outcome variable	Within-sex standardized vs raw attention
Treatment of missing data	Listwise deletion vs multiple imputation
Sample weights	Incorporated vs not incorporated
Covariate set	Original vs expanded
TV age	~1.5 vs ~3
Propensity score analysis, IPTW weighting (384 total)	
Outcome variable	Within-sex standardized vs raw attention
TV cutpoint	20/80, 30/70, 40/60, 50, 60, 70 *(percentiles, 6 levels)*
Sample weights	Incorporated vs not incorporated
Covariate set	Original vs expanded
TV age	~1.5 vs ~3
Treatment effect	ATT vs ATE
Doubly-robust estimation	Doubly robust vs no additional covariate adjustment
Propensity score analysis, Stratification (240 total)	
Outcome variable	Within-sex standardized vs raw attention
TV cutpoint	20/80, 30/70, 40/60, 50, 60, 70 *(percentiles, 6 levels)*
Sample weights	Incorporated vs not incorporated
Covariate set	Original vs expanded
TV age	~1.5 vs ~3
Number of strata	4, 5, 6, 7, 8 *(5 levels)*

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

Table 2
Marginal descriptive statistics for continuous variables

Variable	Valid n	Mean	Std Dev	Min	Max
Age (yrs) when attention was measured	2108	7.75	0.61	6.75	8.75
Annual family income (thousands)	1958	33.42	24.53	0.00	189.92
Attention (raw)	2108	2.64	0.39	1.00	3.00
.Attention within-sex SS	2075	101.25	13.79	83.00	136.00
CES-D Depression score (1992)	2089	46.97	7.87	32.30	79.90
Cognitive stimulation of home age 1-3	1907	97.61	16.15	11.10	148.20
Emotional support of home age 1-3	1765	97.99	16.58	31.60	124.70
Gestational age at birth	1960	-1.41	1.96	-14.00	7.00
Mother's age at birth	2108	28.48	2.62	23.00	36.00
Mother's years of schooling	2095	12.95	2.48	0.00	20.00
Number of children in household	2097	1.64	1.20	0.00	7.00
Partner's years of schooling	1757	13.28	2.70	1.00	20.00
Rosenberg self-esteem score (1987)	2040	45.07	8.40	23.50	59.70
Temperament	1961	2.01	0.69	1.00	5.00
TV hours per day age 1.5	1993	2.23	3.07	0.00	16.00
TV hours per day age 3	2023	3.68	3.12	0.00	16.00

Table 3
Marginal descriptive statistics for categorical variables

Variable	Value	n	Percent
Maternal alcohol use in pregnancy	No	1050	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	1681	79.74%
	Yes	399	18.93%
	(missing)	28	1.33%
Child sex	Female	1034	49.05%
	Male	1074	50.95%
Low birth weight (< 2500g)	No	1812	85.96%
	Yes	138	6.55%
	(missing)	158	7.50%
Health condition that limits school or play	No	1917	90.94%
	Yes	122	5.79%
	(missing)	69	3.27%
Premature birth	No	1744	82.73%
	Yes	216	10.25%
	(missing)	148	7.02%
Child race	Black	572	27.13%
	Hispanic	397	18.83%
	White	1139	54.03%
Maternal smoking in pregnancy	No	1447	68.64%
	Yes	528	25.05%
	(missing)	133	6.31%

Standard metropolitan statistical area (urbanicty)	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%

Table 4
Inverse probability of treatment weighted (IPTW) propensity score model results by TV cutpoint

TV cutpoint percentiles	Non-sig	Sig	Proportion sig
20/80	21	11	0.344
30/70	22	10	0.312
40/60	15	17	0.531
50	12	20	0.625
60	6	26	0.812
70	18	14	0.438

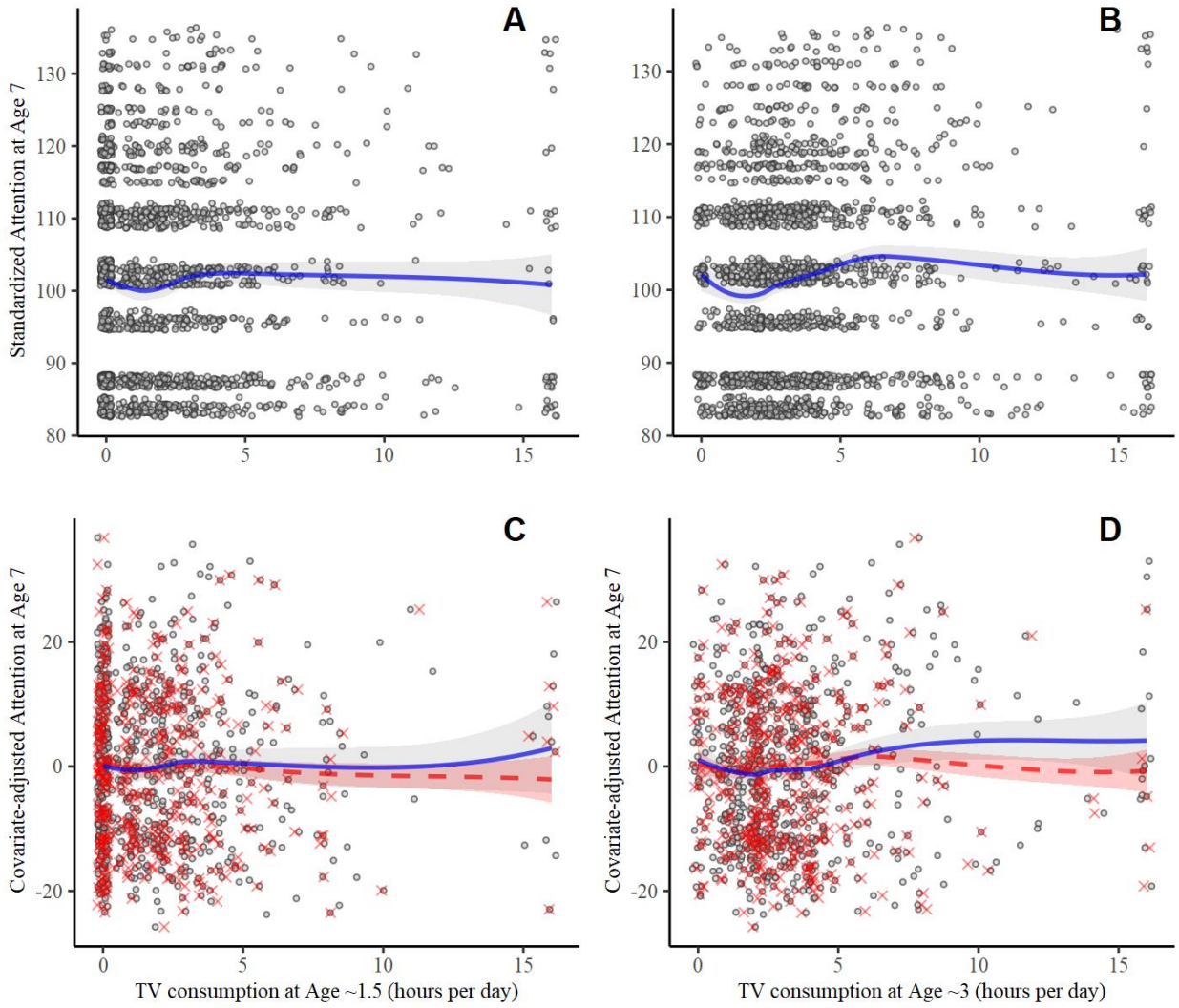
Note: Table included only models measuring TV use at age ~3. When two numbers are given for the cutpoint percentile, this implies that TV use between those percentiles were dropped. So 20/80 means that the low-TV group was defined as the 20th percentile or lower and the high-TV group as the 80th percentile or higher. When a single value is given, it means TV use below that percentile was categorized low-TV, and TV use above it was categorized as high-TV. *Non-sig*: the number of models using the specified attention cutpoint. *Sig*: the number of models that yielded statistical significance.

Table 5
Logistic regression results by attention cutpoint and missing data treatment

Attention cutpoint	Listwise	Non-sig	Sig	Proportion sig	Multiple imputation	Non-sig	Sig	Proportion sig
110		8	0	0.000		4	0	0.000
111		8	0	0.000		4	0	0.000
112		8	0	0.000		4	0	0.000
113		8	0	0.000		4	0	0.000
114		8	0	0.000		4	0	0.000
115		7	1	0.125		4	0	0.000
116		7	1	0.125		4	0	0.000
117		6	2	0.250		4	0	0.000
118		6	2	0.250		4	0	0.000
119		5	3	0.375		4	0	0.000
120		3	5	0.625		2	2	0.500
121		3	5	0.625		2	2	0.500
122		3	5	0.625		2	2	0.500
123		1	7	0.875		2	2	0.500
124		1	7	0.875		2	2	0.500
125		0	8	1.000		2	2	0.500
126		0	8	1.000		2	2	0.500
127		0	8	1.000		2	2	0.500
128		0	8	1.000		3	1	0.250
129		0	8	1.000		3	1	0.250
130		0	8	1.000		3	1	0.250

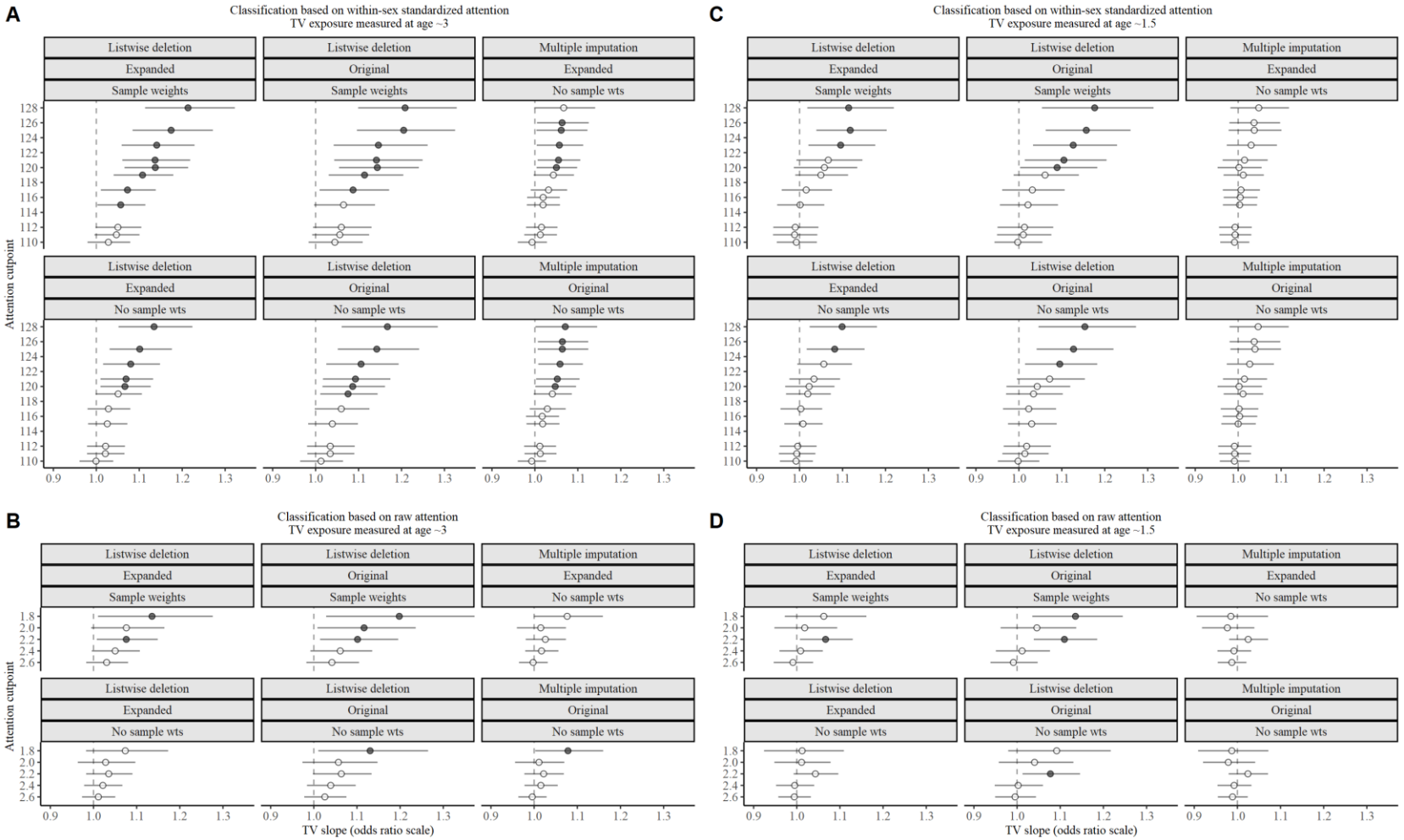
Note: *Non-sig*: the number of models using the specified attention cutpoint and missing data treatment that did not yield statistical significance. *Sig*: the number of models that yielded statistical significance. Cutoffs are given for the within-sex standardized attention scores; percentile-equivalent cutoffs were applied to the raw attention scores.

Figure 1
Scatterplots of early childhood TV use versus standardized within-sex attention score at age 7.



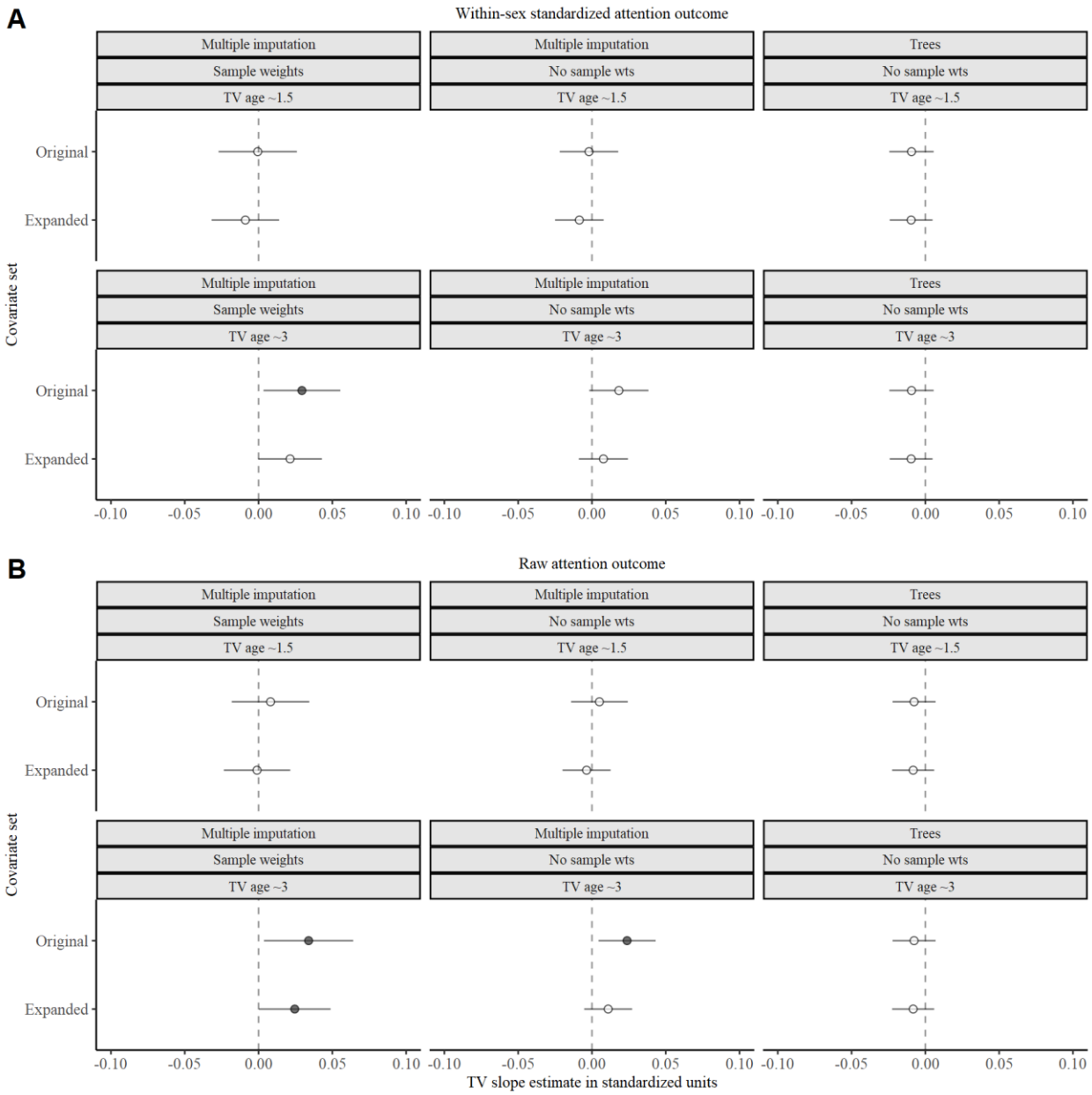
Note: *Left column*: TV measured at age ~1.5. *Right column*: TV measured at age ~3. *Top row*: raw data. *Bottom row*: adjusted (residualized) attention score with effect of covariates removed. Bottom panel: red 'x' points are adjusted based on imputed covariate values. Solid blue smoothing line fits to non-missing data only; red dashed smoothing line fit all data (including imputed values). Point locations are slightly jittered to reduce overplotting

Figure 2
Multiverse I: Logistic regression results summary



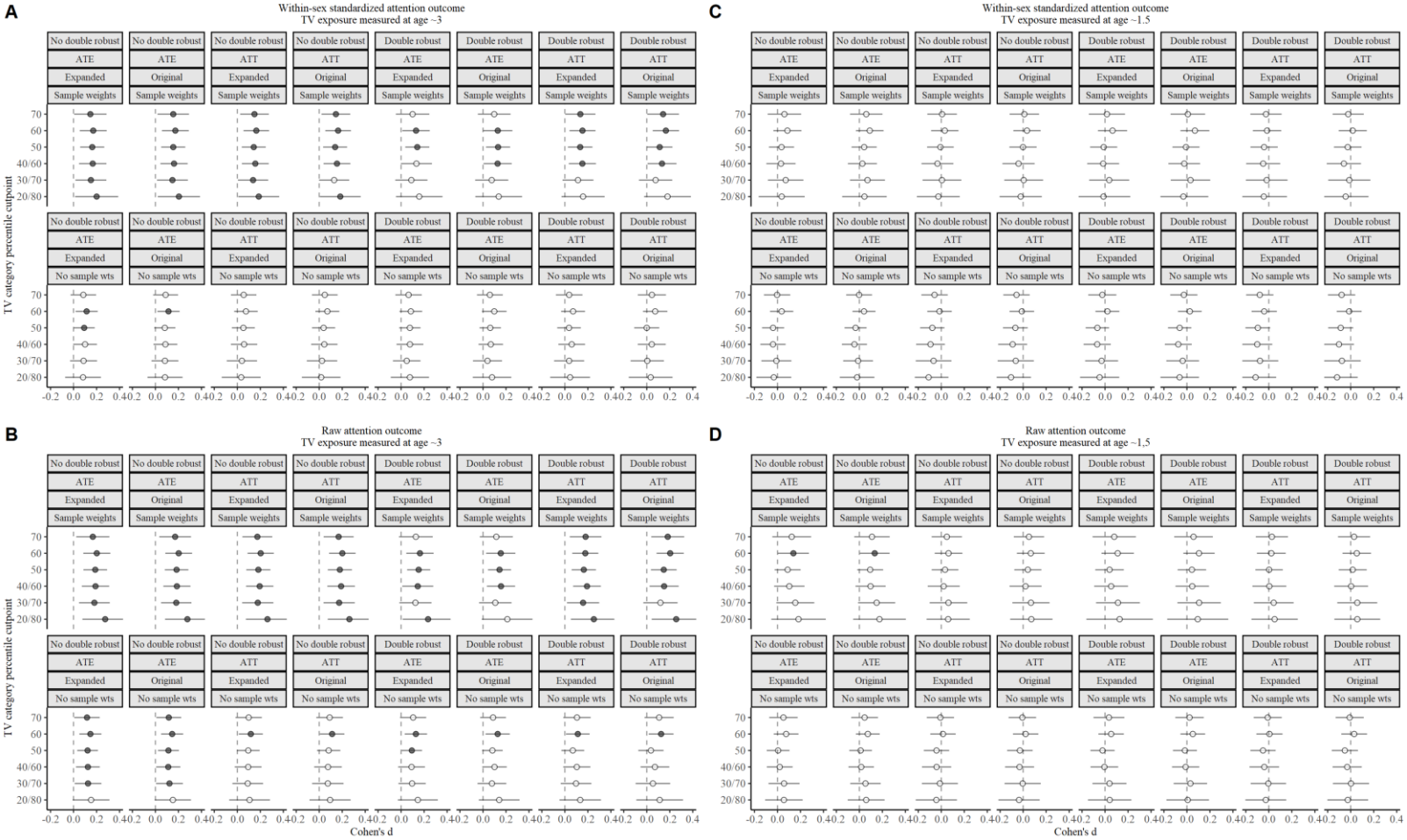
Note: Odds ratio point estimate and 95% CI displayed for each model. Filled circles indicate $p < .05$. *Panel A*: Within-sex standardized attention outcome and TV measured at approximate age 3. *Panel B*: Raw attention outcome and TV measured at approximate age 3. *Panel C*: Within-sex standardized attention outcome and TV measured at approximate age 1.5. *Panel D*: Raw attention outcome and TV measured at approximate age 1.5. Y-axis of each panel shows the cutpoint defining problematic attention. The dashed vertical reference line represents no association (OR=1).

Figure 3A
Multiverse II: Linear regression model results summary



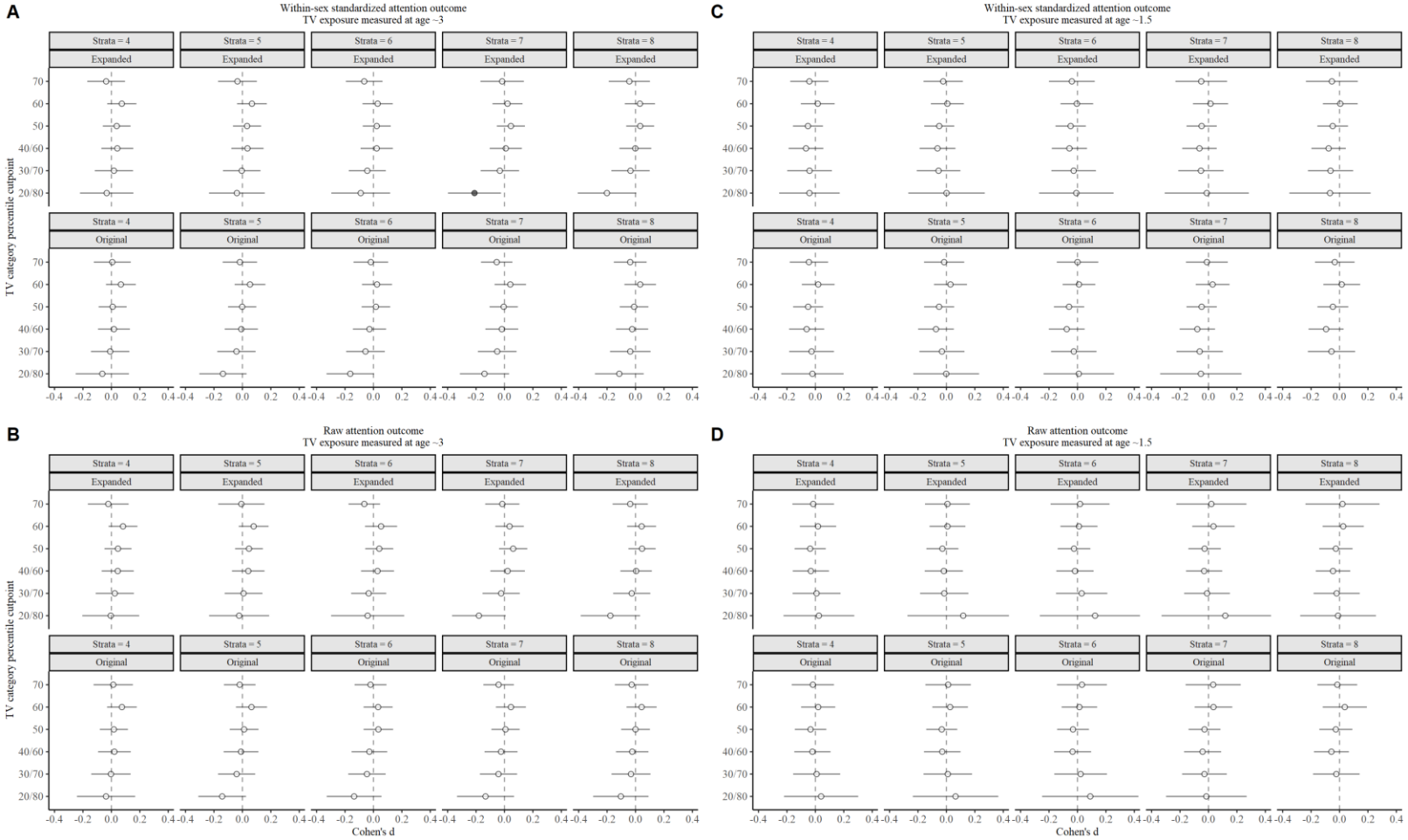
Note: The standardized regression slope for TV and 95% CI displayed for each model. Filled circles indicate $p < .05$. *Panel A*: Within-sex standardized attention outcome. *Panel B*: Raw attention outcome. Y-axis of each panel shows the covariate set. Other model features are listed in the header to each pane. The outcomes are scaled such that higher scores indicate worse attention. The estimates describe the expected in attention, measured in standard deviation units, for a one-hour increase in TV use. The dashed vertical reference line represents no association (b=0).

Figure 3B
Multiverse II: IPTW weighting propensity score analysis results



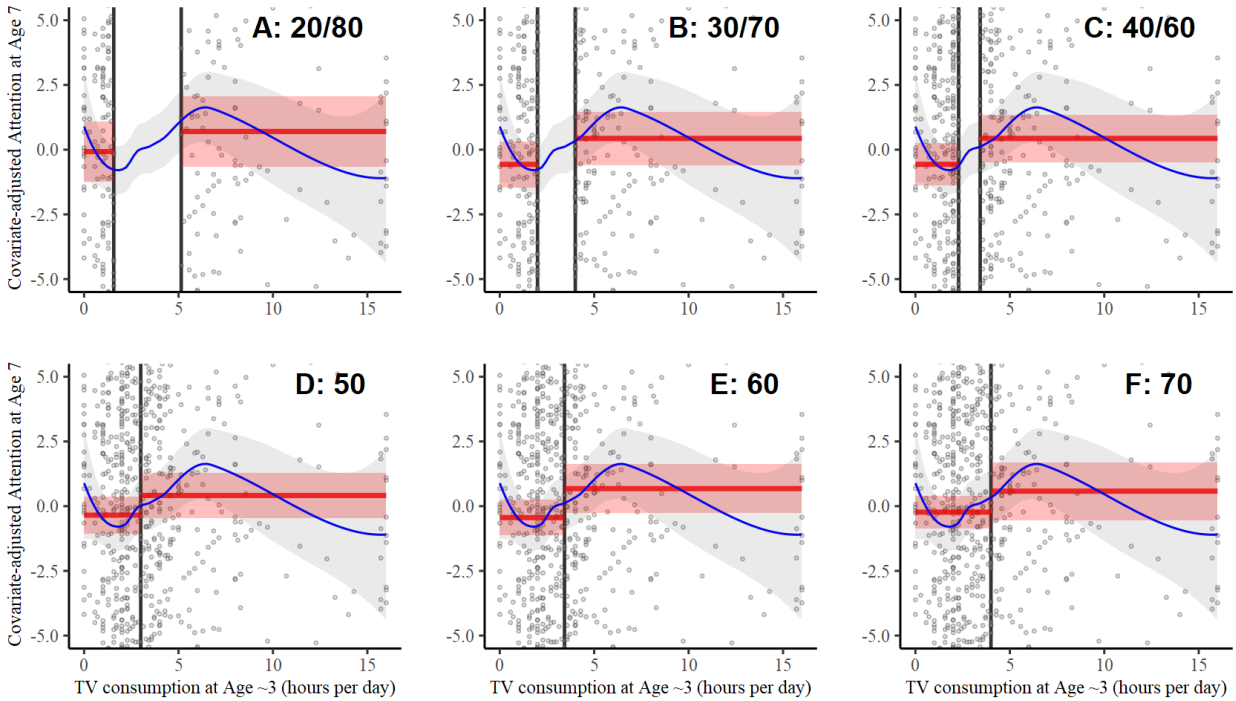
Note: Standardized effect size (Cohen's *d*) estimate and 95% CI presented for each model. Filled circles indicate $p < .05$. *Panel A*: Within-sex standardized attention outcome and TV measured at approximate age 3. *Panel B*: Raw attention outcome and TV measured at approximate age 3. *Panel C*: Within-sex standardized attention outcome and TV measured at approximate age 1.5. *Panel D*: Raw attention outcome and TV measured at approximate age 1.5. Y-axis of each panel shows the percentile cutpoint(s) defining the high- and low-TV groups. Other model features are listed in the header to each pane. ATE = average treatment effect. ATT = average treatment effect for the treated. The outcomes are scaled such that higher scores indicate worse attention. The dashed vertical reference line represents no association (*d*=0).

Figure 3C
Multiverse II: Stratification propensity score analysis results



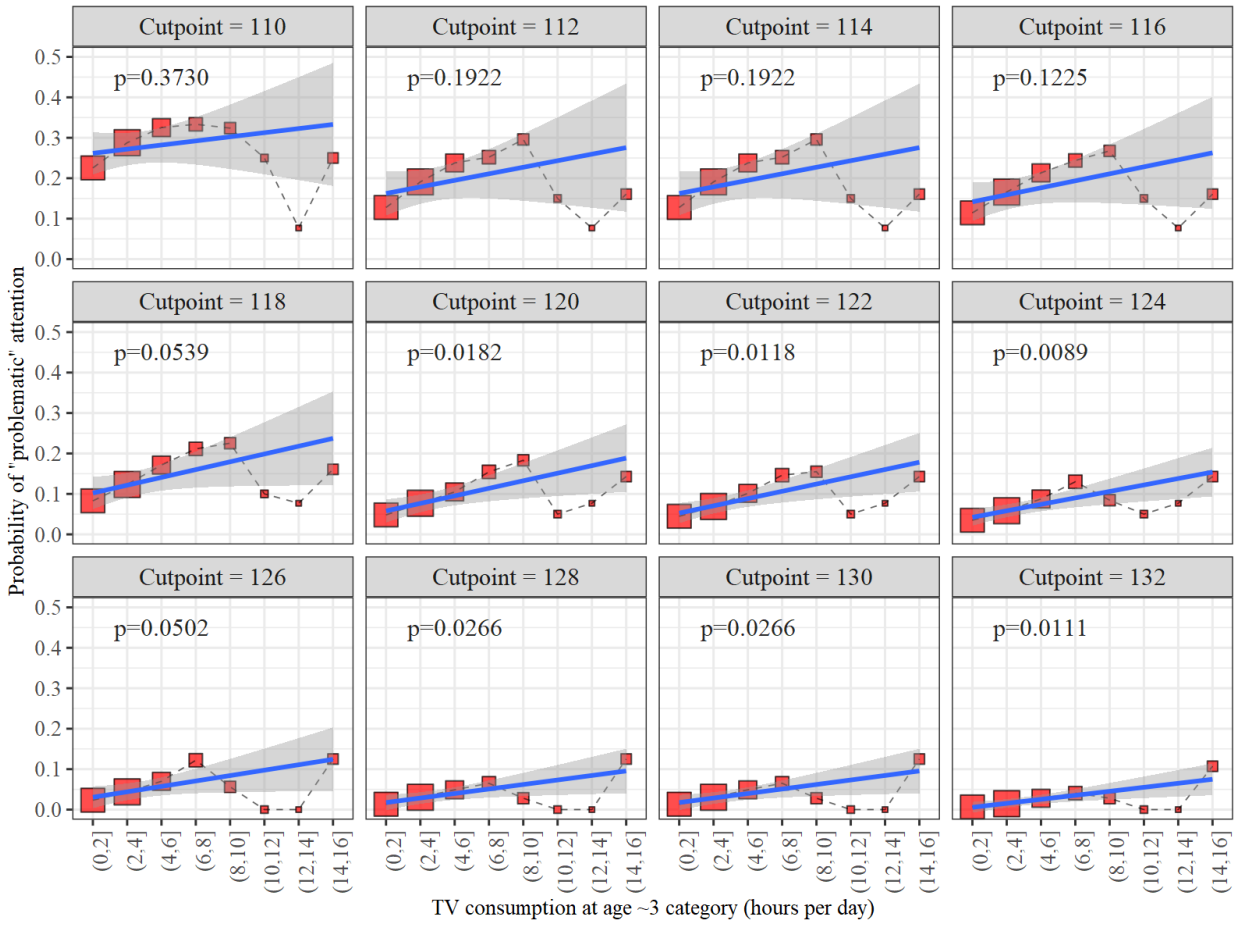
Note: Standardized effect size (Cohen's *d*) estimate and 95% CI presented for each model. Filled circles indicate $p < .05$. *Panel A*: Within-sex standardized attention outcome and TV measured at approximate age 3. *Panel B*: Raw attention outcome and TV measured at approximate age 3. *Panel C*: Within-sex standardized attention outcome and TV measured at approximate age 1.5. *Panel D*: Raw attention outcome and TV measured at approximate age 1.5. Y-axis of each panel shows the percentile cutpoint(s) defining the high- and low-TV groups. Other model features are listed in the header to each pane. The outcomes are scaled such that higher scores indicate worse attention. The dashed vertical reference line represents no association (d=0).

Figure 4
IPTW propensity score model post-mortem



Note: Panels display the zoomed-in residualized attention outcome (e.g., controlling for covariates) versus TV measured at age ~3. Each panel depicts a different set of TV percentile cutpoints for defining the low- and high-TV groups; which are displayed as dark vertical lines and in the label for each panel. The blue curve on each plot is the loess smoother, illustrating the location and scale of the nonlinear feature. The red horizontal lines and their shaded uncertainty regions display the conditional mean attention in the low- and high-TV categories.

Figure 5
Logistic regression post-mortem



Note: Panels display the probability of impaired attention (*y*-axis), as defined by the attention cutpoint displayed on each panel, versus TV measured at age ~3 (*x*-axis). The size of each point is proportional to the number of cases in that level of TV consumption. A fitted regression line (weighted by the sample size within each TV grouping) and its slope *p*-value are depicted on each plot; these represent the performance of the logistic regression model in each situation. As the cutoff rises, the association between TV and the probability of impaired attention seems to increase.

Appendix
Guide for reproducing this analysis using the Docker image

The Docker container is preloaded with the versions of R, RStudio, and all the R packages that were used to perform the analyses reported in this paper.

1) Make a Docker account at http://www.docker.com

2) Log in and download Docker Desktop. If prompted, make sure Docker is set to use Linux containers.

3) Start Docker Desktop, logging in to your account. Under Docker Desktop – Settings – Advanced, make sure that the Docker Engine can use at least 6 GB of memory. Insufficient resources can cause the code to hang. Under Settings – Shared Drives, grant access to one of your local drives so you can copy the generated files out of the container to your local machine.

4) Open a Terminal (Mac / Linux) or Command Prompt (Windows)

5) Type `docker run --rm -e PASSWORD=TV -p 8787:8787 mmcbee/rstudio_tvattention:psychscience`
(If this image isn't found on the local machine, it will be downloaded automatically from Docker Hub)

6) Open a browser tab and navigate to this url: localhost:8787
RStudio will begin running in your browser.

7) Log in to RStudio with username `rstudio`, password `TV`

8) In RStudio, open the file `/Code/analysis.r` in the Files pane

9) Run the code by highlighting it all (Ctrl-A or Cmd-A) and then Ctrl-Enter or Cmd-Enter (Note: it will take several hours to run).

10) Inspect the results in the `/Results` and `/Manuscript/Tables` and `/Manuscript/Figures` folders.

11) End the Docker session by pressing Ctrl-C or in the terminal or command prompt window. On a Mac, this will end the Docker session. On Windows machines you'll need to determine either the Container ID or the Name of the session by typing `docker ps`

The Container Id is 12-character string such as `b1971e3eea21`. It will be different each time you run the container. Next, type `docker stop CONTAINERID`, for example, `docker stop b1971e3eea21`. Alternatively, you can refer to the container by its name, which will be a combination of random words such as `priceless_galois` or `competent_ellis`. The container's name and id are shown by `docker ps`

Copying files from the Docker container to your local machine

After the analysis script is finished running, you will likely want to copy the files to your local computer for examination. If you stop the Docker container, you will have to run the analysis again to recreate all the files.

1) Determine the Container Id for the Docker session by typing `docker ps`. On a Mac you will need to open a new Terminal window, as pressing Ctrl-C in the active Terminal window will end the Docker session. On a Windows machine, pressing Ctrl-C will allow you to enter additional commands in the active Command Prompt window without disturbing the Docker session.

2) Change the directory in your Command Prompt / Terminal session to the local directory to which you want to copy the files with the `cd` command. For example, `cd "c:\Users\Matt\Documents\TVAttention"` (Windows) or `cd "/home/Users/Matt/Documents"` (Mac)

3) Copy the files by typing `docker cp CONTAINERID:home/rstudio .` For example, `docker cp b1971e3eea21:home/rstudio .` (You can substitute the container name for the Container id).