Supporting Information

Participants. Participants came from two independent cohorts of twin subjects enrolled around the age of 11. Older-cohort MTFS subjects were recruited from all twins born during the relevant birth years who met minimal inclusion criteria and who could be located [11]. The sample is thus a population-based random sample. The sampling frame for the ES was designed to enrich the sample for adolescents at high risk for substance abuse through a combination of a screening interview completed by mothers of prospective twin participants and a randomly drawn group of twins [14]. The majority of ES subjects for the present investigation were randomly sampled, however (as were all MTFS subjects). Not all ES participants could be assessed for the age-20 follow-up due to a lapse in grant funding.

Alcohol use measures. We derived a measure of drinking by combining responses to questions about four different aspects of drinking: frequency of use, typical quantity consumed, density of use (maximum number of drinks in a 24-hr period) and misuse (drinking to intoxication). The distribution of responses to several of these questions was somewhat sparse and right-skewed. We created ordinal scales from each by collapsing across similar responses as described in **Table S1**.

Table S1: *Converting raw responses to ordinal scales of different aspects of drinking.*

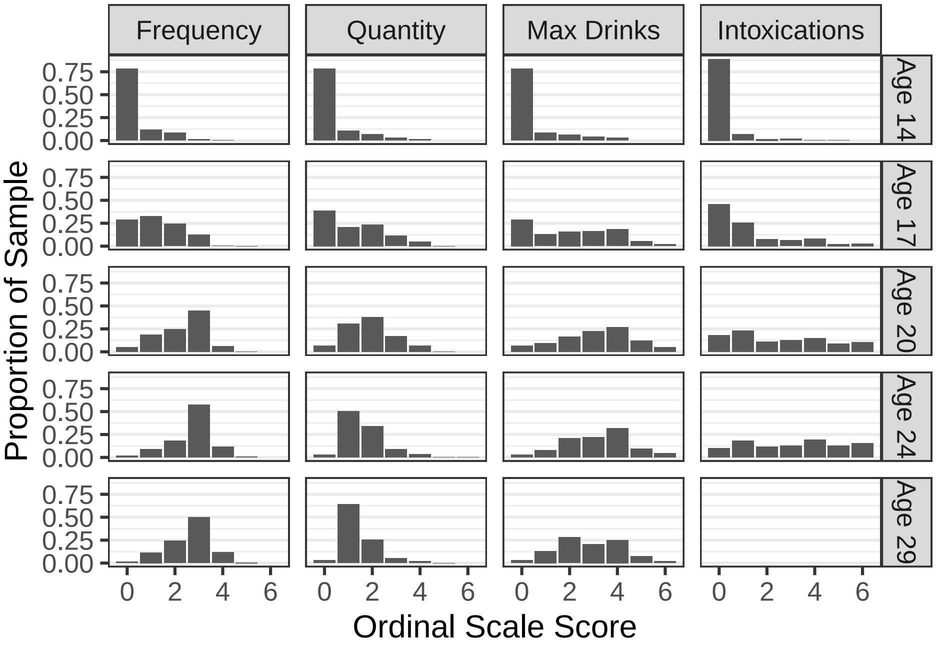
Scale Score	Drinking Frequency	Typical Quantity	Maximum Quantity	Intoxications
0	Never or not in the past year	0	0	0
1	Less than once a month	1-3	1-3	1-5
2	1-3 times per month	4-6	4-6	6-10
3	1-4 times per week	7-10	7-10	11-20
4	Every day or nearly every day	11-20	11-20	21-50
5	2 or more times a day	21-29	21-29	51-149
6	NA	30+	30+	150+

Note: The table refers to the manner in which we converted raw scores to ordinal scales (“Scale Score”) from responses to the relevant questions on the Substance Abuse Module (SAM) [19]. Questions and response options on the computerized questionnaire were similar although not identical. The ordinal scale for frequency of drinking consisted only of 5 levels. “Typical Quantity” refers to the number of drinks typically consumed when one drinks. “Maximum Quantity” is the maximum number of drinks consumed in a single 24-hr period. “Intoxications” refers to the number of times drinking to the point of being intoxicated. At the age-17 assessment, the questions about frequency of drinking and typical quantity consumed concerned the previous year, whereas the questions about the maximum amount consumed and the number of times intoxicated concerned the individual’s lifetime. At subsequent follow-ups, the first three questions concerned the time since the previous assessment. Number of times intoxicated always concerned the participant’s lifetime.

Figure S1 shows the distribution of responses on the ordinal scale derived as in **Table S1** across the different assessment waves. Participants weren’t asked how many times they had become intoxicated at the age-29 assessment as a result of the need to shorten the length of participants’ visits to the university. Although at some assessments the SAM [19] and computerized inventory [8] were both administered, for ease of presentation we used responses to questions from the SAM for the age-17 assessment and subsequent follow-ups.

Propensity score indicators and estimation. The cotwin-control design cannot control for unshared confounders. In order to capture any such influences, we derived a propensity score estimating each subject’s propensity for levels of lifetime alcohol use. Indicators were drawn from multiple domains assessed as part of the comprehensive age-11 intake assessment in both cohorts. Because there was slight variation in the assessment protocol between the two cohorts, we selected items that were common to both. Items were identified on a rational basis and those that demonstrated correlations with cumulative alcohol use of .10

Figure S1: *Drinking item frequencies by assessment wave.*



or greater in absolute value were retained. Nearly all were characteristics that might differ between twins.

Table S2: *Propensity score indicators.*

Indicator	Informant/Source	Instrument	Reliability	Pct Missing
Socialization (n = 6)	Teacher	Teacher ratings	0.87	6.4
Socialization (n = 4)	Self	DBI items	0.5	6.2
Boldness (n = 8)	Teacher	Teacher ratings	0.8	6.3
Withdrawn Behavior (n = 4)	Teacher	Teacher ratings	0.77	7.0
Attitude Toward School (n = 6)	Teacher	Teacher ratings	0.91	30.2
Academic Problems (n = 6)	Mother	Parent interview	0.83	13.7
Academic Motivation (n = 6)	Mother	Parent interview	0.83	11.7
Externalizing Symptoms (n = 35)	Mother	DICA-R	0.71–0.81	0.0
Externalizing Behavior (n = 49)	Teacher	Teacher ratings	0.82	6.7
Conflict with Parents (n = 12)	Self	Parent Environment Questionnaire	0.84	7.7
Deviant Peers (n = 9)	Self	Computerized questionnaire	0.78	39.1
Life Stress (n = 18)	Self	Life Events Inventory	0.42	3.4
Family Occupation	Mother	Hollingshead scale	NA	2.8
Maximum Alcohol Consumption	Parents	Substance Abuse Module	NA	0.0
Birthweight	Mother/Birth Records	NA	NA	0.0

Note: N in parentheses after each indicator is the number of items making up that indicator. If missing, the indicator consists of a single item. The Reliability column provides Cronbach's α measure of internal consistency. Teacher ratings consisted of ratings of 1–4 teachers nominated by the child and his or her mother. Teachers rated the individual twins on items adapted from the Conners Teacher Rating Scale [5] and the Rutter Child Scale B [22]. Other items covered DSM criteria for externalizing psychopathology, academic progress, personality trait ratings, and peer group characteristics. We took the mean rating across teachers. DBI is the 36-item Delinquent Behavior Inventory [7], administered by computer. Items for the Socialization indicator comprise part of a measure reflecting a premorbid liability for substance abuse, as are the teacher-rated Socialization and Boldness items [9]. DICA-R is the revised Diagnostic Interview for Children and Adolescents–Parent version [18, 26]. Symptoms of conduct disorder, oppositional defiant disorder and ADHD were summed and log-transformed. Reliability estimates (kappa coefficients) are for presence of each disorder. The Parent Environment Questionnaire was developed by the MCTFR [6]. Responses to questions about the relationship with each parent were averaged together. Life Events Interview items used here were the child's responses to binary, yes/no questions assessing living instability, personal losses, family financial difficulties, parental conflict and interpersonal problems. Responses to the 18 items were summed. Occupational status was assessed using the Hollingshead scale [10]. We used the maximum value for the two parents. The Substance Abuse Module of the Composite International Diagnostic Interview (CIDI) [19, 20] was modified by the MCTFR to include questions about quantity and frequency of drinking. The mother's and father's maximum number of drinks consumed in a single 24-hr period consistently predicts adolescent offspring substance use and abuse [15, 16].

Not every indicator was available for every subject due to incomplete data. In addition, two questions were not asked of the MTFS males. (See Table **S2** for the percentage of missing data for each indicator.) We therefore imputed missing data, using the panImpute wrapper to the R package `pan`, as implemented in `mitml`. `pan` is designed for clustered samples. It assumes a multivariate normal distribution for the measures to be imputed. We created 50 imputation sets, using the combined cohorts in order to leverage data available for ES males in particular in imputing missing scores for MTFS males. To ensure independence of the 50 sets, we specified 50,000 burn-in iterations and 5,000 iterations between imputation sets. The propensity score indicators predicted cumulative exposure equally well in the two cohorts, with a multiple R averaged over imputation sets of .40 and .38 for ES and MTFS, respectively.

The distribution of baseline covariates is independent of exposure when conditioned on the true propensity score [1]. Propensity scores are thus "balancing scores" in that they approximately balance the level of covariates influencing treatment or exposure between groups. The generalized propensity score (GPS) balances the level of all covariates across levels of a continuous measure of exposure [12, 13]. We used the CBPS (Covariate Balancing Propensity Score) approach in the `CBPS` package in R, which estimates a propensity score and optimizes balance simultaneously. Propensity scores were subsequently used as inverse probability of treatment weights (IPTW) [2] in reanalyses of significant within-pair effects. We derived the twin deviation from his or her respective twin-pair mean weight, in order to retain the desired interpretation of the within-pair effect in the CTC design [23, 24]. We added a constant (the absolute value of the minimum weight) to all such weights in order to avoid negative weights, and trimmed weights at the 1^{th} and 99^{th} quantiles to avoid numerical instability due to extreme values (including 0). Weights were normalized so as to sum to the number of subjects in the sample. Results of IPTW-weighted analyses using these trimmed weights were combined across imputation sets as prescribed by Rubin and colleagues [3, 21] to yield propensity score-adjusted estimates of the within-pair effect. Significant effects that become nonsignificant after such weighting would suggest that the exposure–performance association was at least partially due to unshared confounding. **Figure S2** illustrates the effectiveness of the CBPS procedure in balancing covariates (propensity score indicators) across levels of alcohol use. Whereas correlations between raw indicators and the propensity score ranged in absolute value to greater than $r = 0.30$, all were virtually zero after weighting.

Results

Cotwin-control model. The CTC design decomposes each individual's drinking score into orthogonal components: the twin-pair mean and each individual's deviation from the mean. Ignoring covariates, this takes the form algebraically $\hat{Y}_{ij} = b_0 + b_W(X_{ij} - \bar{X}_{.j}) + b_B\bar{X}_{.j}$, where $\hat{Y}_{ij}$ is the expected performance score, X_{ij} is the drinking score for individual i in twin pair j, $\bar{X}_{.j}$ is the mean drinking score for twin pair j and b_0 is the model intercept.

Adjusting for IQ. As described in the text, we included IQ as a covariate in follow-up analyses of individual-

Figure S2: *Correlations between propensity score indicators and alcohol use before and after balancing.*

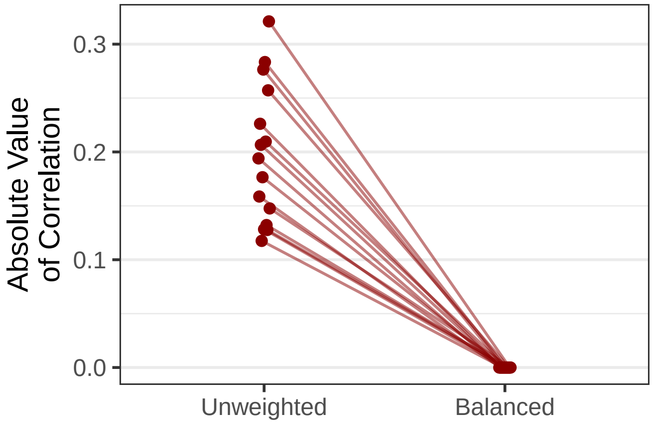
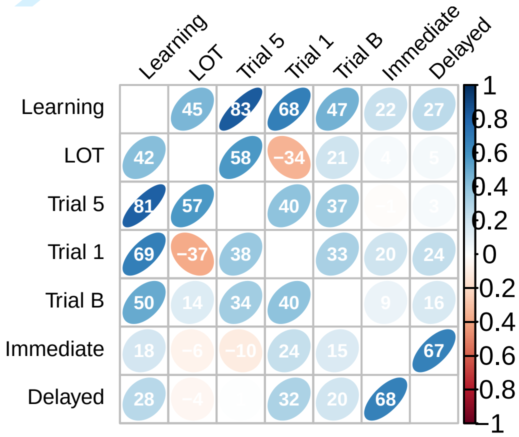


Figure S3: *Correlations among performance measures in the two age cohorts.*



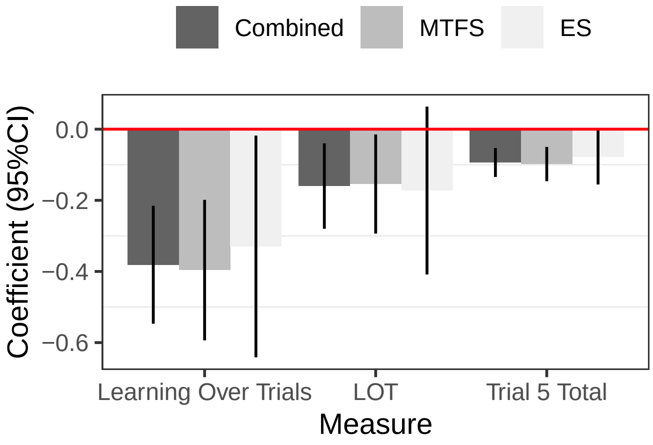
	Learning	LOT	Trial 5	Trial 1	Trial B	Immediate	Delayed
Learning		45	83	68	47	22	27
LOT	42		58	−34	21	4	5
Trial 5	81	57		40	37	−1	3
Trial 1	69	−37	38		33	20	24
Trial B	50	14	34	40		9	16
Immediate	18	−6	−10	24	15		67
Delayed	28	−4		32	20	68	

Note: "Learning" refers to the measure of overall (total) learning, while "Immediate" and "Delayed" refer to immediate and delayed retention, respectively (see main text for details).

level associations between cumulative alcohol use and task performance on the RAVLT as well as for within-pair estimates in the cotwin-control design. IQ was assessed at the age-11 intake assessment by means of the Wechsler Intelligence Scales for Children – Revised (WISC-R) [25]. Despite being measured many years earlier, IQ was significantly associated with all measures of performance except proactive interference (the Trial B—-Trial 1 difference score), with t-statistics for the other measures ranging from 2.76 to 11.47 and all p-values $\leq$ .0001, despite having been assessed either nine or 18 years earlier, on average.

Correlation among performance measures. The relationship among trials was similar in the two age cohorts. Cronbach's alpha was identical ($\alpha = 0.88$). **Figure S3** graphically illustrates the pattern of correlation among the different performance measures separately for the two age cohorts. Correlations for the ES cohort are in the lower triangle and those for the MTFS cohort are in the upper triangle. The ellipse in each cell represents the magnitude of the correlation, while its color indicates the direction. The more concentrated the ellipse is along its main axis, the larger the correlation. As the figure shows, the pattern of correlations is very similar for the two cohorts, indicating that the measures cohere in much the same way.

Figure S4: *Consistency of phenotypic associations across cohorts.*



Consistency of results across cohorts. As indicated in the manuscript, cohort by alcohol use interaction effects were not significant for any performance measure. In **Figure S4** we plot the magnitude of parameter estimates and 95% confidence intervals around them separately for the two cohorts as well as for the combined sample. Measures showing a significant association with alcohol use in primary analyses, as described in the text, were included. The figure illustrates the similarity in findings across cohorts for the three measures of learning. Confidence intervals for the ES cohort were wider than for the MTFS cohort, but point estimates are similar.

Are associations specific to alcohol use? **Table S3** presents parameter estimates and test statistics assessing associations between cumulative cannabis use and task performance, as described in the manuscript. The left-hand columns of the table present cannabis use–task performance associations adjusted only for covariates of no interest (cohort, sex and zygosity), whereas the right-hand columns present the same associations but with propensity scores used as inverse probability to treat weights (IPTW). Propensity score-adjusted results in the right-hand columns of **Table S3** reflect cannabis use–task performance associations when the propensity for alcohol exposure is held approximately constant in the sample. That these were all nonsignificant indicates that the significant associations betwewen cannabis use and measures of learning and attention listed in the left-hand columns of the table were likely attributable to propensity to drink rather than to direct effects of cannabis use.

Table S3: *Phenotypic associations between cumulative cannabis use and RAVLT task performance, with and without adjusting for propensity for alcohol use.*

	Unadjusted Cannabis Effects				Propensity Score-Adjusted Effects						
	Estimate	SE	t-statistic	p-value	Estimate	SE	t-statistic	df	p-value	RIV	FMI
Overall Learning	-0.817	0.293	-2.79	0.005	-0.649	0.425	-1.53	46.8	0.133	0.028	0.027
Learning Over Trials	-0.226	0.198	-1.14	0.254	-0.218	0.296	-0.74	46.7	0.465	0.029	0.028
Trial 5 Total	-0.225	0.075	-3.01	0.003	-0.182	0.114	-1.60	46.8	0.117	0.028	0.028
Trial 1 Total	-0.118	0.053	-2.24	0.025	-0.086	0.059	-1.46	47.0	0.150	0.022	0.022
Trial B Total	-0.153	0.057	-2.71	0.007	-0.084	0.080	-1.05	47.3	0.300	0.018	0.018
Trial B–Trial 1	-0.035	0.062	-0.57	0.570	0.003	0.084	0.03	46.9	0.976	0.026	0.025
Immediate Retention	-0.007	0.056	-0.13	0.896	0.048	0.084	0.57	46.7	0.573	0.030	0.029
Delayed Retention	0.033	0.061	0.54	0.593	0.046	0.083	0.55	46.5	0.583	0.033	0.032

Note: Estimate is the parameter estimate for cumulative cannabis use and SE its associated standard error, obtained with the cluster-robust sandwich estimator in svyglm. All parameter estimates are adjusted for any effects of cohort, sex and zygosity. Unadjusted estimates and associated statistics are for raw individual-level associations, whereas estimates in the right-hand columns are propensity score-adjusted estimates and associated statistics. Propensity score indicators were all from the age-11 assessment (see **Table S2**). Missing values were imputed 50 times, a propensity score estimated for each imputation set, and phenotypic analyses were conducted on each set using IPTW weighting. Results of the 50 sets of IPTW-weighted analyses were combined using "Rubin's rules" [21]. df are corrected as recommended by Barnard and Rubin [3]. RIV is the relative increase in variance due to nonresponse and FMI is the fraction of missing information, both of which quantify the influence of missing data on a parameter's sampling variance.

In addition, we conducted a mediation analysis at the individual level (Level 1 in hierarchical linear models). Twin deviation scores for cannabis use as well as alcohol use served as predictors. Cannabis use deviation scores were not significantly associated with any of the task performance measures. Results are displayed in the left-hand columns of **Table S4**. Thus, twin differences in cannabis use did not predict task performance when adjusted for twin differences in alcohol use. By contrast, alcohol use deviation scores *were* significantly associated with all measures of learning and attention, as indicated in the right-hand columns of **Table S4**.

Table S4: *Cannabis use–performance associations assessed via a mediation analysis at Level 1.*

	Cannabis Within-Pair Effects				Alcohol Within-Pair Effects			
	Estimate	SE	t-statistic	p-value	Estimate	SE	t-statistic	p-value
Overall Learning	-0.480	0.506	-0.95	0.342	-0.560	0.186	-3.02	0.003
Learning Over Trials	-0.312	0.472	-0.66	0.509	-0.407	0.170	-2.39	0.017
Trial 5 Total	-0.153	0.142	-1.07	0.284	-0.141	0.051	-2.77	0.006
Trial 1 Total	-0.034	0.113	-0.30	0.766	-0.031	0.041	-0.75	0.456
Trial B Total	-0.144	0.116	-1.24	0.215	-0.044	0.042	-1.05	0.296
Trial B–Trial 1	-0.111	0.149	-0.75	0.455	-0.013	0.050	-0.26	0.794
Immediate Retention	0.135	0.138	0.98	0.329	-0.013	0.048	-0.28	0.781
Delayed Retention	-0.130	0.139	-0.94	0.348	0.042	0.049	0.85	0.395

Note: Estimate is the parameter estimate for the within-pair effect (the twin-difference effect represented by the individual twin's deviation from their respective twin-pair mean). SE is its associated standard error, obtained with the cluster-robust sandwich estimator in svyglm. Estimates and their associated statistics for the cannabis use within-pair effect are in the left-hand columns, while estimates of the alcohol use within-pair effect on the same outcome measures are in the right-hand columns. These were adjusted for effects of each other as well as cohort, sex and zygosity.

Discussion

Our failure to obtain evidence of significant associations between drinking and performance on the two recall trials of the RAVLT, as is often reported, may be due in part to differences in the definition of recall. Like other studies, we obtained sizeable associations between cumulative alcohol use and the sheer number of words recalled on recall trials (not shown). However, the dependent measure in our regression models consisted of retention, defined as the number of words recalled minus the number of words learned by the final learning trial. A recent longitudinal study using a similar definition of retention to separate recall from learning obtained similar results regarding binge drinkers relative to a comparison group [17]. Taken together, these results suggest that the poorer performance of adolescents or young adults on recall trials may reflect a general deficit rather than a deficit of recall per se. However, it may also be that poor recall is a specific correlate of clinically significant and heavy alcohol consumption [4].

Supplement References

1. Austin PC. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behav Res 2011;46:119–151.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research 2011;46:399–424.
3. Barnard J and Rubin DB. Miscellanea. small-sample degrees of freedom with multiple imputation. Biometrika 1999;86:948–955.
4. Brown SA, Tapert SF, Granholm E, and Delis DC. Neurocognitive functioning of adolescents: effects of protracted alcohol use. Alcohol Clin Exp Res 2000;24:164–71.
5. Conners C. Conner's Teacher Rating Scale. Iowa City, IA: University of Iowa, 1969.
6. Elkins IJ, McGue M, and Iacono WG. Genetic and environmental influences on parent-son relationships: evidence for increasing genetic influence during adolescence. Developmental Psychology 1997;33:351–63.
7. Gibson H. Self-report delinquency among school boys and their attitudes to police. British Journal of Social and Clinical Psychology 1967;20:303–315.
8. Han C, McGue M, and Iacono WG. Lifetime tobacco, alcohol and other substance use in adolescent minnesota twins: univariate and multivariate behavioural genetic analyses. Addiction 1999;94:981–993.
9. Hicks BM, Iacono WG, and McGue M. Index of the transmissible common liability to addiction: heritability and prospective associations with substance abuse and related outcomes. Drug and Alcohol Dependence 2012;123:S18–S23.
10. Hollingshead AB. Two factor index of social position. New Haven, CT: Author, 1957.
11. Iacono WG, Carlson SR, Taylor J, Elkins IJ, and McGue M. Behavioral disinhibition and the development of substance use disorders: findings from the Minnesota Twin Family Study. Development and Psychopathology 1999;11:869–900.
12. Imai K and Van Dyk DA. Causal inference with general treatment regimes: generalizing the propensity score. Journal of the American Statistical Association 2004;99:854–866.
13. Imbens GW. The role of the propensity score in estimating dose-response functions. Biometrika 2000;87:706–710.
14. Keyes MA, Malone SM, Elkins IJ, Legrand LN, McGue M, and Iacono WG. The Enrichment Study of the Minnesota Twin Family Study: increasing the yield of twin families at high risk for externalizing psychopathology. Twin research and human genetics: the official journal of the International Society for Twin Studies 2009;12:489.
15. Malone SM, Iacono WG, and McGue M. Drinks of the father: father's maximum number of drinks consumed predicts externalizing disorders, substance use, and substance use disorders in preadolescent and adolescent offspring. Alcoholism: Clinical and Experimental Research 2002;26:1823–1832.

16. Malone SM, McGue M, and Iacono WG. Mothers' maximum drinks ever consumed in 24 hours predicts mental health problems in adolescent offspring. Journal of child psychology and psychiatry 2010;51:1067–1075.

17. Nguyen-Louie TT, Tracas A, Squeglia LM, Matt GE, Eberson-Shumate S, and Tapert SF. Learning and memory in adolescent moderate, binge, and extreme-binge drinkers. Alcohol Clin Exp Res 2016;40:1895–904.

18. Reich W. Diagnostic Interview for Children and Adolescents (DICA). Journal of the American Academy of Child and Adolescent Psychiatry 2000;39:59–66.

19. Robins LN, Babor TF, and Cottler LB. Composite International Diagnostic Interview: Expanded Substance Abuse Module. St. Louis: Authors, 1987.

20. Robins LN, Wing J, Wittchen HU, et al. The Composite International Diagnostic Interview. an epidemiologic instrument suitable for use in conjunction with different diagnostic systems and in different cultures. Archives of General Psychiatry 1988;45:1069–77.

21. Rubin DB. Multiple imputation for nonresponse in surveys. John Wiley & Sons, 2004.

22. Rutter M. A children's behavior questionnaire for completion by teachers: preliminary findings. Journal of Child Psychology and Psychiatry 1967;8:1–11.

23. Saunders GRB, McGue M, and Malone SM. Sibling comparison designs: addressing confounding bias with inclusion of measured confounders. Twin Res Hum Genet 2019:1–7.

24. Sjölander A, Frisell T, and Öberg S. Causal interpretation of between-within models for twin research. Epidemiologic Methods 2012;1:217–237.

25. Wechsler D. Manual for the Wechsler Intelligence Scale for Children–Revised. San Antonio, TX: The Psychological Corporation, 1981.

26. Welner Z, Reich W, Herjanic B, Jung K, and Amado H. Reliability, validity, and parent-child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). Journal of the Academy of Child and Adolescent Psychiatry 1987;26:649–653.

R Version and R Packages

1. Fong C, Ratkovic M, and Imai K. CBPS: Covariate Balancing Propensity Score. R package version 0.14. 2017.
2. Gohel D. flextable: Functions for Tabular Reporting. R package version 0.5.10. 2020.
3. Grund S, Robitzsch A, and Luedtke O. mitml: Tools for Multiple Imputation in Multilevel Modeling. R package version 0.3-7. 2019.
4. Lumley T. Analysis of complex survey samples. Journal of Statistical Software 2004;9. R package verson 2.2:1–19.
5. Lumley T. survey: analysis of complex survey samples. R package version 3.35-1. 2019.
6. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2017.
7. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
8. Xie Y. Dynamic Documents with R and knitr. 2nd. ISBN 978-1498716963. Boca Raton, Florida: Chapman and Hall/CRC, 2015.
9. Zhao JH and Schafer JL. pan: Multiple imputation for multivariate panel or clustered data. R package version 1.4. 2016.