Chapter 1

Research Governance, Methodological Framework and Reproducibility

Teyfide Tecel Hatipoğlu*

**Northern Cyprus Prime Minister's Anti-Drug Commission, Server Somuncuoğlu Street, No:10, Flat:1, Köşklüçiftlik / Nicosia, 9001, North Cyprus*

DOI: 10.70020/BI.20260803.1

1.1. Ethical Governance

This study was conducted in accordance with the ethical standards of the European School Survey Project on Alcohol and Other Drugs (ESPAD) and the principles outlined in the Declaration of Helsinki (World Medical Association, 2025), which encompass the requirements for voluntary participation, protection of participant welfare, and maintenance of confidentiality. Participation in the survey was strictly voluntary, and students were informed at the outset that they could withdraw at any stage without encountering any adverse consequences. No personally identifiable information was collected at any stage of the research.

The data collection process was conducted in accordance with the General Data Protection Regulation (GDPR, EU Regulation No. 2016/679) and relevant national data protection laws.

Prior to the commencement of field implementation, official administrative permissions were secured through a coordinated multi-tiered governance chain involving the highest relevant state authorities in Northern

Cyprus. Initial strategic authorization was issued by the Prime Ministry, which formally referred the research infrastructure to the relevant administrative bodies. Consequently, institutional approval was granted by the Ministry of National Education through the Office of the Undersecretary, permitting the survey implementation across public secondary educational institutions under the purview of both the General Secondary Education Directorate and the Vocational and Technical Education Directorate. Furthermore, the Northern Cyprus Prime Minister Anti Drug Commission evaluated and approved the survey protocol, confirming its alignment with national youth prevention frameworks and public health priorities. At the final operational stage, the school principals of all participating educational institutions reviewed the ministerial directives and provided the necessary school level institutional consent to facilitate data collection within their respective campuses.

The survey was administered electronically via Google Forms, allowing participants to complete the questionnaire independently on digital devices. To uphold the voluntary and anonymous nature of the study, the digital interface was configured to ensure privacy, with no personal identifiers or tracking metadata being collected. This digital administration format was selected to provide a confidential environment for participants, thereby maximizing the validity of self-reported data in sensitive research domains and eliminating the risk of observer bias or interpersonal influence during the response process.

1.2. Casual Framework

Causal inference from observational data requires an explicit representation of the assumed data-generating process. In this study, Directed Acyclic Graphs (DAGs) were utilized as the primary tool to encode causal assumptions prior to analysis. A Directed Acyclic Graph was constructed for each core outcome to identify (a) the set of variables required for confounding adjustment, (b) variables that function as mediators and should therefore not be conditioned upon in total-effect models, and (c) variables that act as effect-

measure modifiers (moderators) and are entered as interaction terms in pre-specified models.

The variables in this study are classified into five structural roles. Block A comprises demographic variables, including age, gender, school type, region and family socioeconomic status. These are regarded as distal determinants and potential confounders for all subsequent relationships. Block B comprises family-level variables, including family support, parental rule-setting, and parental knowledge of the adolescent's whereabouts, representing proximal family influences that may act as either confounders or mediators on the pathways from demographics to substance use. Block C school variables, such as truancy and academic performance, occupy an intermediate position, potentially mediating the effects of family factors while also functioning as direct risk indicators. Block D peer substance use variables represent domain-specific social influences and, consistent with social learning theory, are treated as proximal determinants of individual use behavior. Finally, Block E perceptual variables -perceived risk and perceived availability- are entered in the final step of the hierarchical models as the most proximal cognitive determinants.

Spurious associations between peer substance use and individual outcomes arising through common environmental factors (e.g., neighborhood context, school culture) represent the primary source of residual confounding in this study. Adjustment for school type and region across all models partially addresses this concern; however, unmeasured neighborhood-level factors remain an acknowledged limitation. Gender is hypothesized to moderate the associations between parental monitoring and substance use, as well as between peer exposure and substance use. These interaction terms were pre-specified and are reported throughout all primary analytical chapters. The assumption that all relevant confounders within each analytical block are measured and adjusted for is an unavoidable simplifying assumption in observational survey research and is acknowledged as an untestable limitation.

1.3. Hypothesis Hierarchy

This study adheres to a pre-specified hypothesis framework designed to distinguish confirmatory inferences from exploratory ones. Four primary hypotheses (H1-H4) were formulated prior to data access; while confirmatory in intent but not prospectively registered, these hypotheses map directly onto the predictor blocks entered into all primary multivariable models and are tested using pre-specified model formulations. Results pertaining to H1-H4 are subjected to Benjamini-Hochberg False Discovery Rate (BH-FDR) correction in sections where multiple outcome groups are tested simultaneously to control for Type I error inflation.

H1 (Parental Monitoring) posits that parental monitoring - operationalized through the rule-setting and parental knowledge subscales- are inversely associated with tobacco use (*Y_VAPE30*), binge drinking (*Y_BINGE30*), cannabis use (*Y_CAN12*), and polydrug use (*Y_POLY*), reflecting the protective role of family-level supervision.

H2 (Peer Exposure) assumes that domain-specific peer substance use are positively associated with all core outcomes, consistent with social learning and peer socialization mechanisms.

H3 (Risk Perception and Availability) assumes that perceived risk are inversely associated with substance use outcomes, while perceived availability are positively associated, reflecting the dual perceptual determinants of adolescent initiation and maintenance of use.

H4 (Gender Moderation) assumes that gender moderates the associations between parental monitoring and substance use, as well as between peer exposure and substance use, such that effect magnitudes differ between male and female students.

Secondary hypotheses cover the relationships between school related variables, specifically truancy and academic performance, and their associations with substance use outcomes. These are tested within the same hierarchical framework but are categorized as nonconfirmatory. Exploratory

analyses, such as elastic net variable selection, latent class analysis, and population attributable fraction estimation, are explicitly labeled in the relevant sections and are not used to draw confirmatory conclusions. Post hoc observations arising from pattern examination are reported with appropriate epistemic caution to ensure the distinction between hypothesis testing and hypothesis generating results remains auditable.

1.4. Pre-Analysis Registration and Deviation Framework

The analytical protocol of this study was established in full detail prior to data access. The study design and structured analytical pipelines were formally submitted to the Prime Ministry Anti-Drug Commission and the Ministry of National Education prior to field access. This protocol encompasses the confirmatory hypothesis framework, primary and secondary model formulations for each analytical chapter, pre-specified interaction terms, the BH-FDR correction strategy, and the criteria for detecting model misspecification. While this pre-specification is consistent with the principle of prospective registration promoted by open science frameworks, no formal registration was filed in a public repository for this study.

Deviations from the pre-specified protocol were unavoidable in several chapters and were transparently documented at the point of occurrence within each analytical chapter, rather than consolidated here. This decision reflects the principle that protocol deviations are best described in immediate proximity to the analyses they affect, ensuring that the reader can evaluate each deviation in its appropriate methodological context. In each case, the deviation is classified as attributable to instrument design, sample characteristics, or data availability, and is clearly distinguished from analytical failure. All post-hoc analytical decisions made after data access are labelled as exploratory and are not reported as confirmatory results.

Across all analytical chapters, it was not possible to implement the pre-specified multilevel random-intercept structure because individual identifiers

for schools were unavailable. Fixed-effects models, including school type and region as covariates, were adopted as the primary analytical approach. This deviation is documented in each chapter, and its impact on standard error estimation is discussed.

1.5. Model Selection Logic

Model selection in this study was determined by the nature of the dependent variable, the clustering structure of the data, and the sample size available for each outcome. The pre-specified primary framework was generalized linear mixed models (GLMM) featuring a random intercept for schools to account for the within-school clustering of student responses. Since individual school identifiers were unavailable in the Northern Cyprus dataset, this pre-specified structure was not implemented. Fixed-effects logistic regression, including school type and region as covariates, was utilized as the primary alternative throughout the study.

For binary outcomes -including *Y_VAPE30*, *Y_BINGE30*, *Y_INT30*, *Y_ALC30*, *Y_RXMIS*, and *Y_GAMB*- standard logistic regression with a logit link function was utilized. For ordinal outcomes with an ordered categorical structure -*Y_BFREQ* (binge drinking frequency) and Y_GAMBFREQ (gambling frequency)- cumulative logistic (proportional odds) regression was utilized; the proportional odds assumption was verified in all cases using the poTest procedure, and violations are reported with explanatory notes where present.

For the continuous gaming composite score (*Y_GAME_score*), standard linear regression was applied, given the approximately normal distribution of the constructed composite. For the pre-specified mediation analysis in Chapter 6, structural equation modeling (SEM) with bootstrap-corrected standard errors was utilized.

For the primary models in each analytical chapter with a sufficient number of cases, Bayesian replication was performed using the brms package.

1.6. Data Transparency and Open Science

The dataset analyzed in this book was derived from the Turkish Republic of Northern Cyprus ESPAD 2025 wave, administered in accordance with the ESPAD 2019 Master Questionnaire protocol and containing items adapted at the national level as permitted by ESPAD guidelines. Variable definitions, response coding schemes, scale construction rules, and all derivation decisions are documented in full in the codebook, which serves as the primary reference for researchers wishing to validate or extend the analyses herein.

All analytical choices made after data access -including decisions necessitated by unexpected data characteristics such as the near-zero outcome prevalence in Chapters 7 and 9 and the singular fit behavior of regional random intercepts in Chapter 8- were explicitly defined as post-hoc decisions and distinguished from the pre-specified analytical program documented in Chapter 0. The analysis log maintained throughout the project records the date, rationale, and consequences of each such decision.

The complete analysis syntax and all R scripts utilized for generating the results reported in this book are available from the corresponding author upon reasonable request. A complete session log documenting the R version, package versions, and system environment at the time of analysis is maintained. This log will be provided alongside the code upon request to ensure full computational reproducibility.

1.7. Reproducible Analysis Workflow

All analyses reported in this book were implemented via scripted code using R software. A single master script reproduces all tables and figures—from raw data to final outputs—without the need for manual intervention or interactive editing. The workflow consists of five sequential stages: (1) raw data import and quality checks, (2) variable recoding and scale construction,

(3) measurement validation using exploratory factor analysis, (4) descriptive analyses, and (5) multivariable modeling and post-estimation.

Each stage was implemented as a separate and annotated module; analytical decisions and protocol deviations were recorded in an analysis log maintained throughout the project.

The reproducibility procedures adopted in this study reflect a commitment to the principles of open and verifiable science. By ensuring that each analytical result is traceable from a single documented input through a transparent sequence of programmatic steps to a specific output, this study design minimizes the risk of analytical errors commonly associated with ad-hoc, interactive, or undocumented workflows. This approach also facilitates the independent replication and extension of the analyses by other researchers working in similar adolescent survey settings. Reporting was guided by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement where applicable.

References

R Core Team. (2025). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

World Medical Association (2025). World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Participants. *JAMA*, *333*(1), 71–74. https://doi.org/10.1001/jama.2024.21972