CHI'26

HANYANG UNIVERSITY

CHUNGBUK NATIONAL UNIVERSITY

LAPS:
Automating Hypothesis-Driven Statistical Analysis of Public Survey Using Large Language Models






Jaehoon Kim* Dayoung Jeong* Beejin Son Hansung Kim Bogoan Kim Kyungsik Han

** Both authors contributed equally to this research*

HANYANG HUMAN CENTERED COMPUTING LAB

CHI'26

HANYANG UNIVERSITY

CHUNGBUK NATIONAL UNIVERSITY

LAPS:
Automating Hypothesis-Driven Statistical Analysis of Public Survey Using Large Language Models






Jaehoon Kim* Dayoung Jeong* Beejin Son Hansung Kim Bogoan Kim Kyungsik Han

** Both authors contributed equally to this research*

HANYANG HUMAN CENTERED COMPUTING LAB

Public Survey Data in Social Science

Understanding Social Dynamics

Essential resources for understanding individual lives and social trends across domains

Key Asset for Social Scientists

Widely utilized to derive critical insights for public policy and academic research

How Social Scientists Analyze Survey Data

1. **Theory-Grounded Hypotheses:** Formulating research hypotheses based on theoretical foundations and previous studies

2. **Operationalization:** Translating abstract concepts into measurable survey variables (reflecting the researcher's domain knowledge)

3. **Statistical Modeling:** Applying appropriate statistical methods to test the relationships between selected variables

Challenges & Needs

Challenge 1

High structural complexity in public surveys causes cognitive overload

Need 1

Support **to identify suitable variable candidates** across scattered items and **explore alternative operationalizations**

Challenge 2

Traditional tools leave **key decision conditions opaque**, making reproducible planning difficult

Need 2

A structured framework **to organize complex methodological conditions**

Challenge 3

General-purpose LLMs produce expansive, unpredictable outputs that drift beyond the intended analytical scope

Need 3

Bounded, evidence-based decision-making tightly **aligned with social science workflows**

Design Goals

Design Goal 1

Provides researchers with **candidate variables** and reasoning, and enables them to **review and modify**

Design Goal 2

Provides researchers with **structured analysis plan** and enables them to **iteratively improve** them

Design Goal 3

Constrain functionality to a **domain-bounded workflow** to ensure a predictable and trustworthy analytical assistant

Need 1

Support to **identify suitable variable candidates** across scattered items and **explore alternative operationalizations**

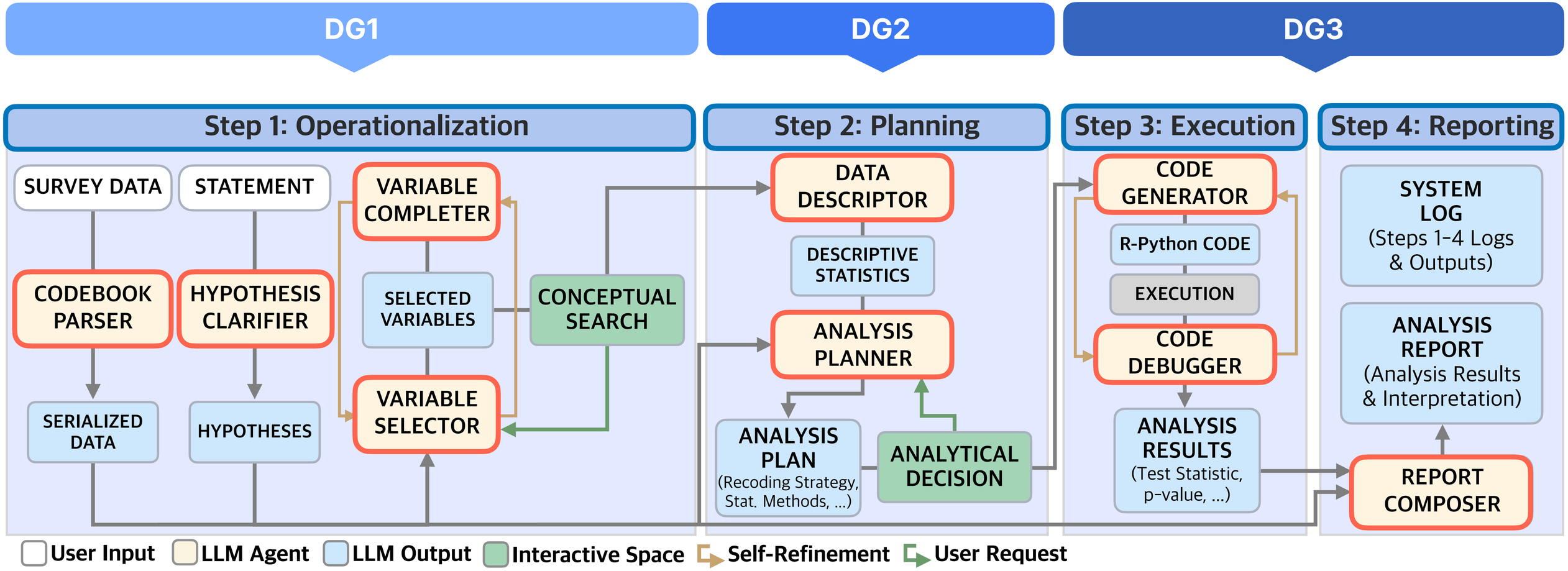
Need 2

A structured framework **to organize complex methodological conditions**

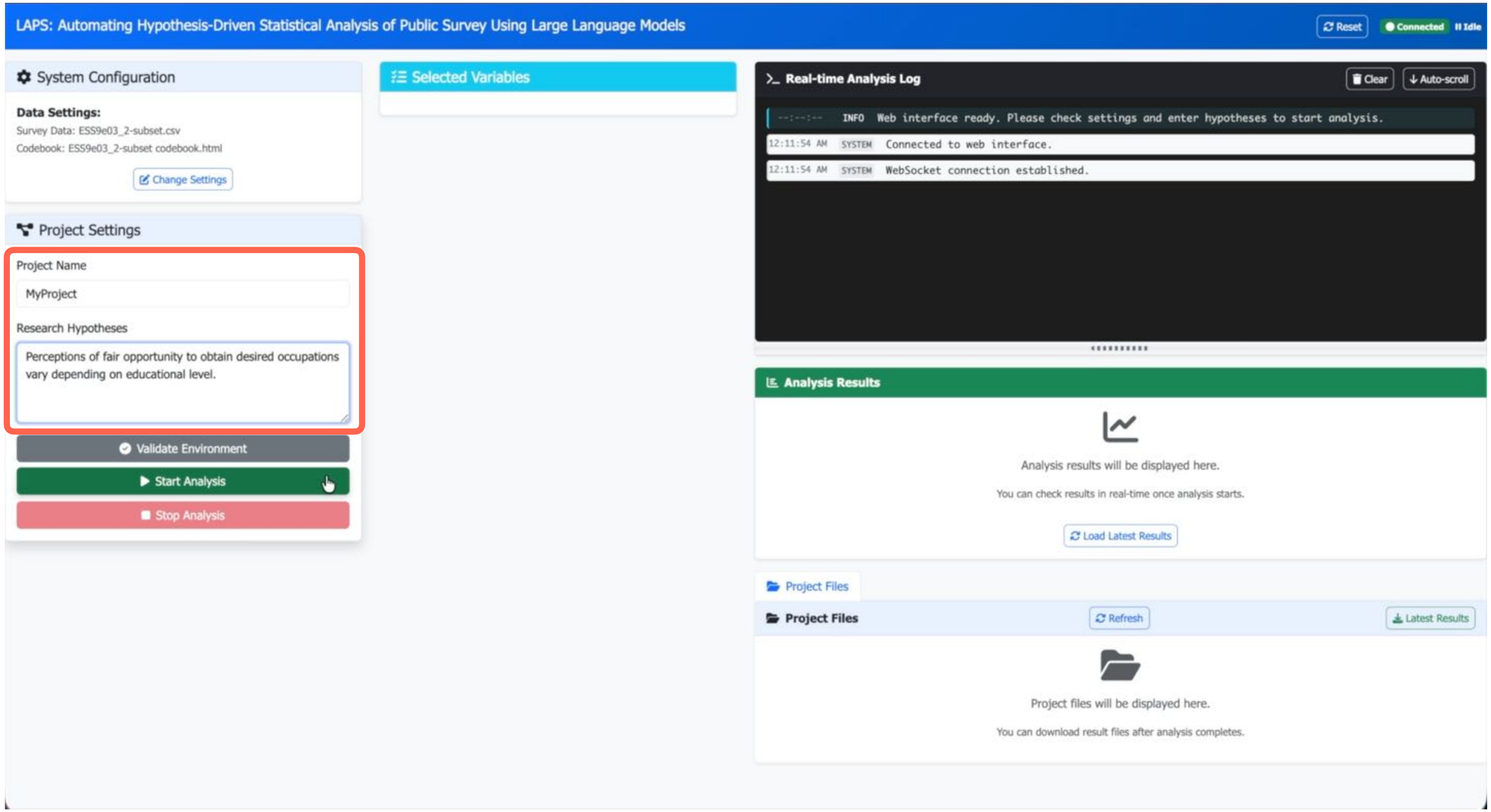
Need 3

Bounded, evidence-based decision-making tightly **aligned with social science workflows**

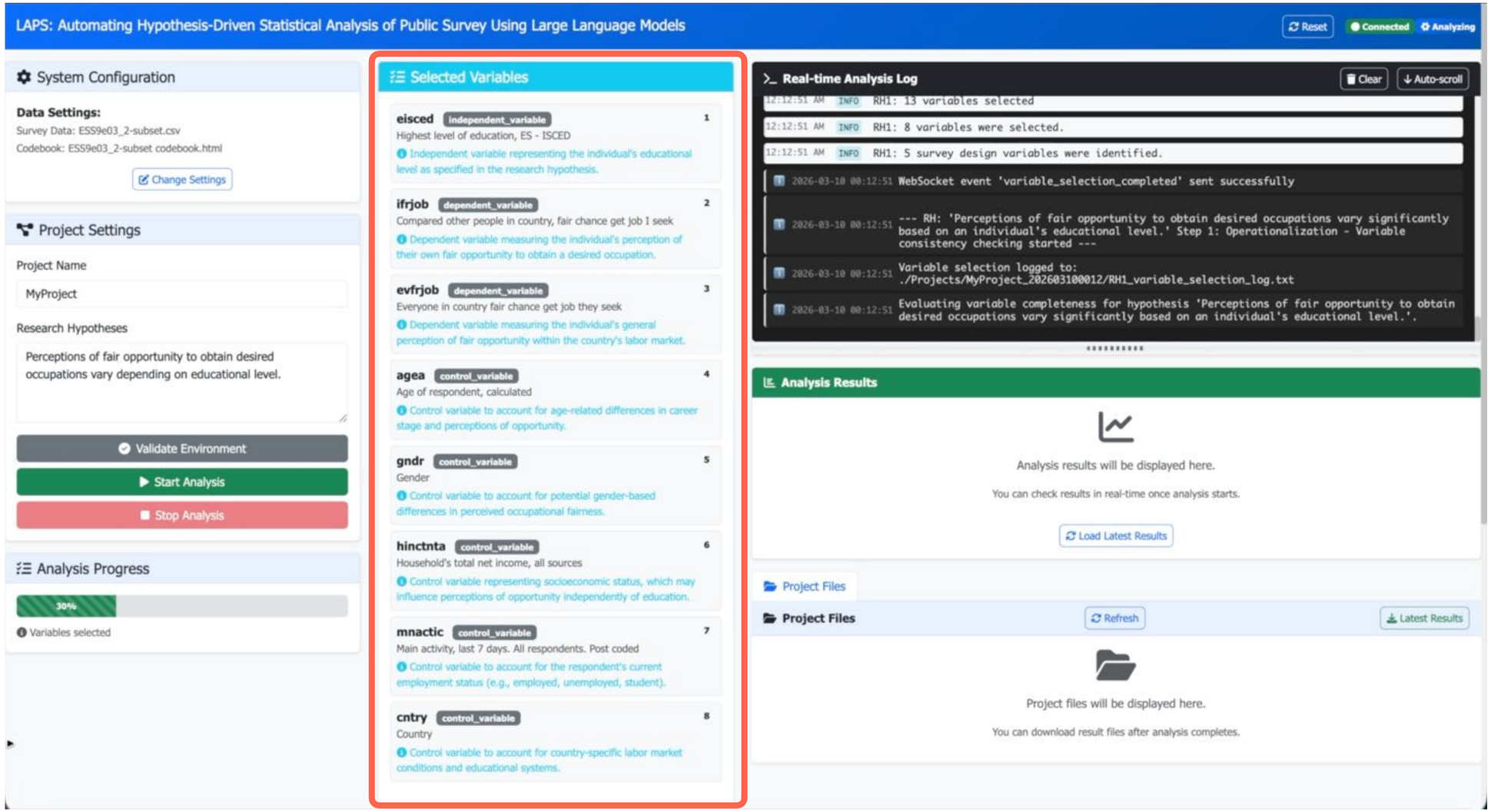
Operational Flow of LAPS



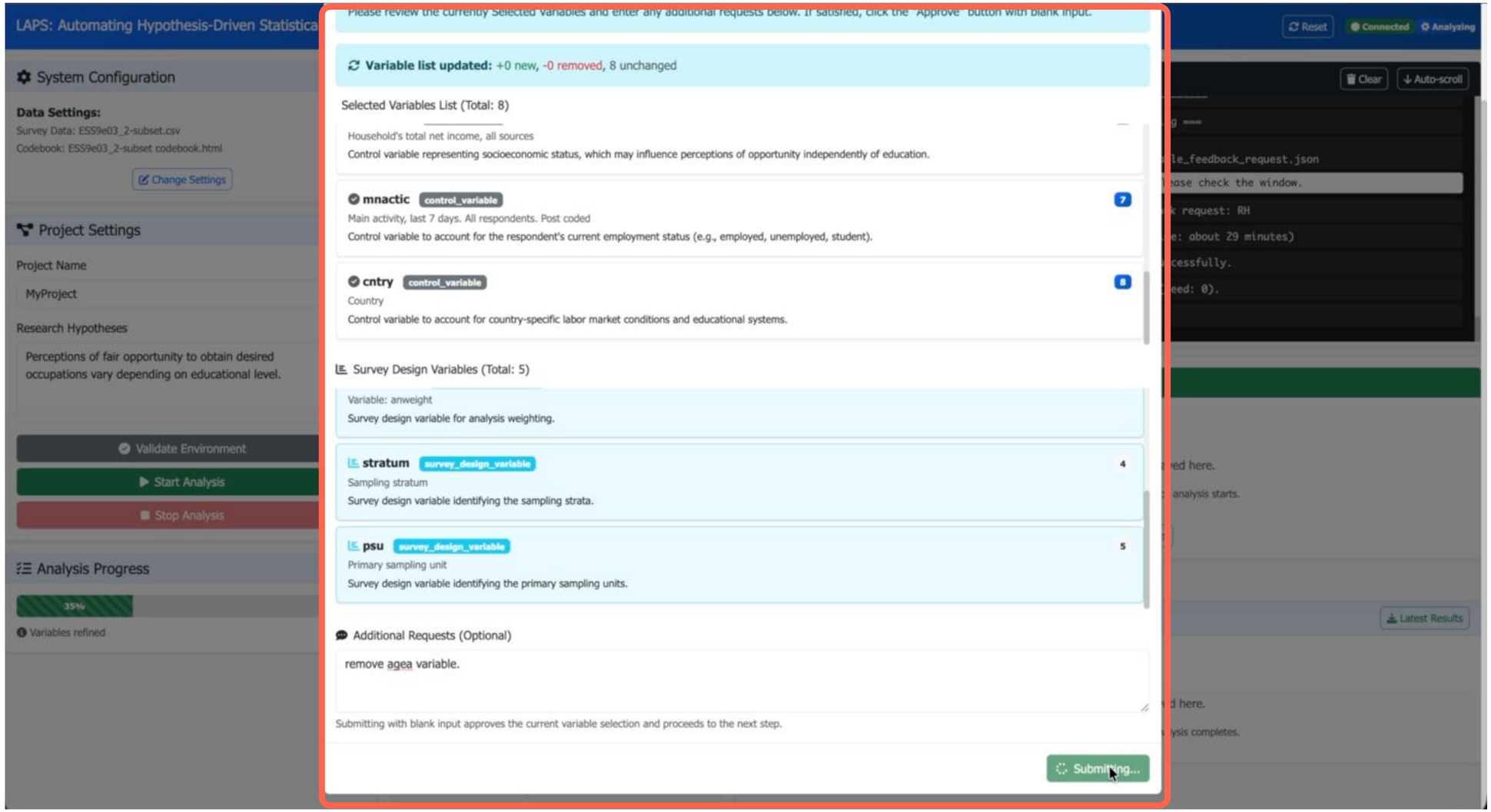
Step 1: Operationalization



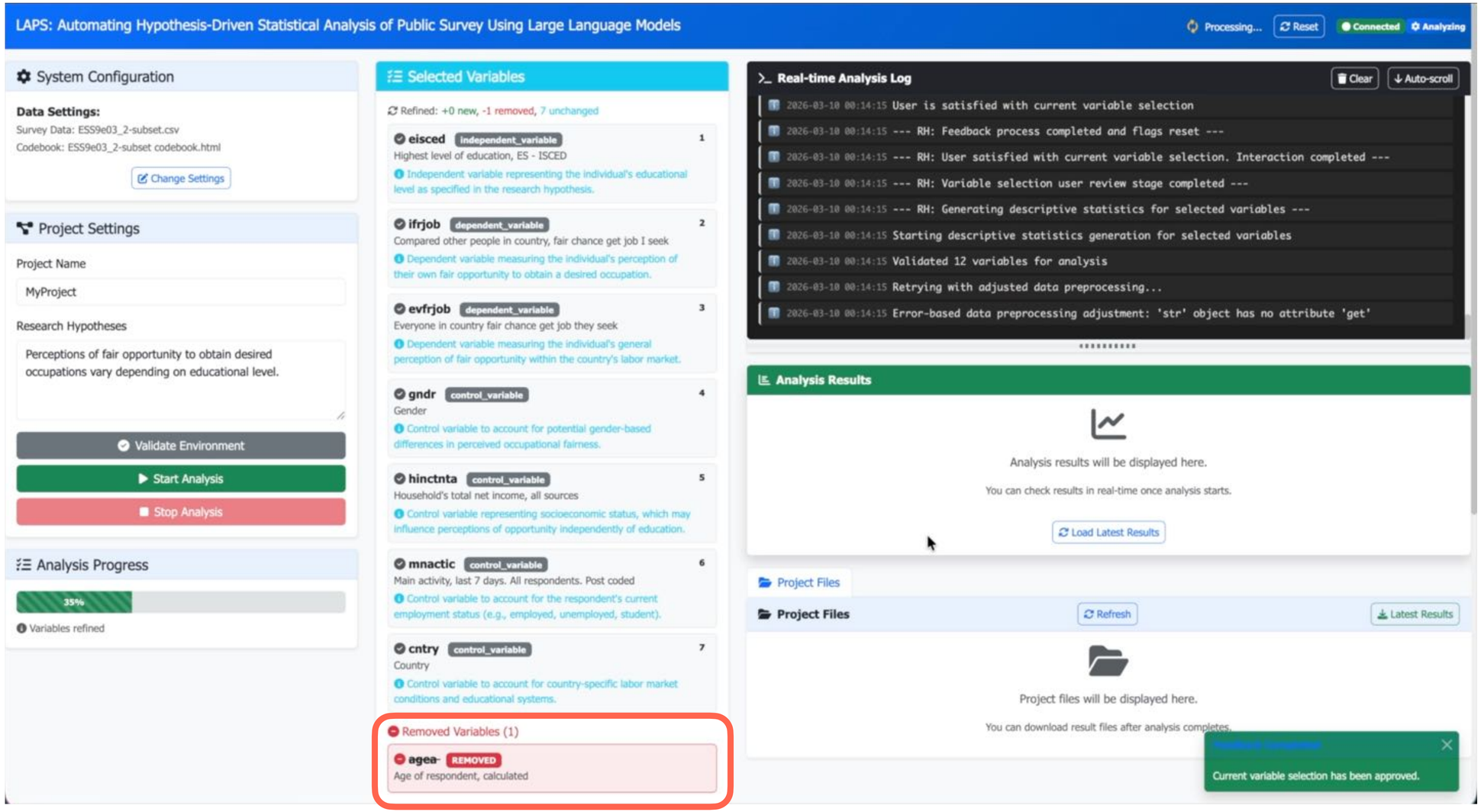
Step 1: Operationalization



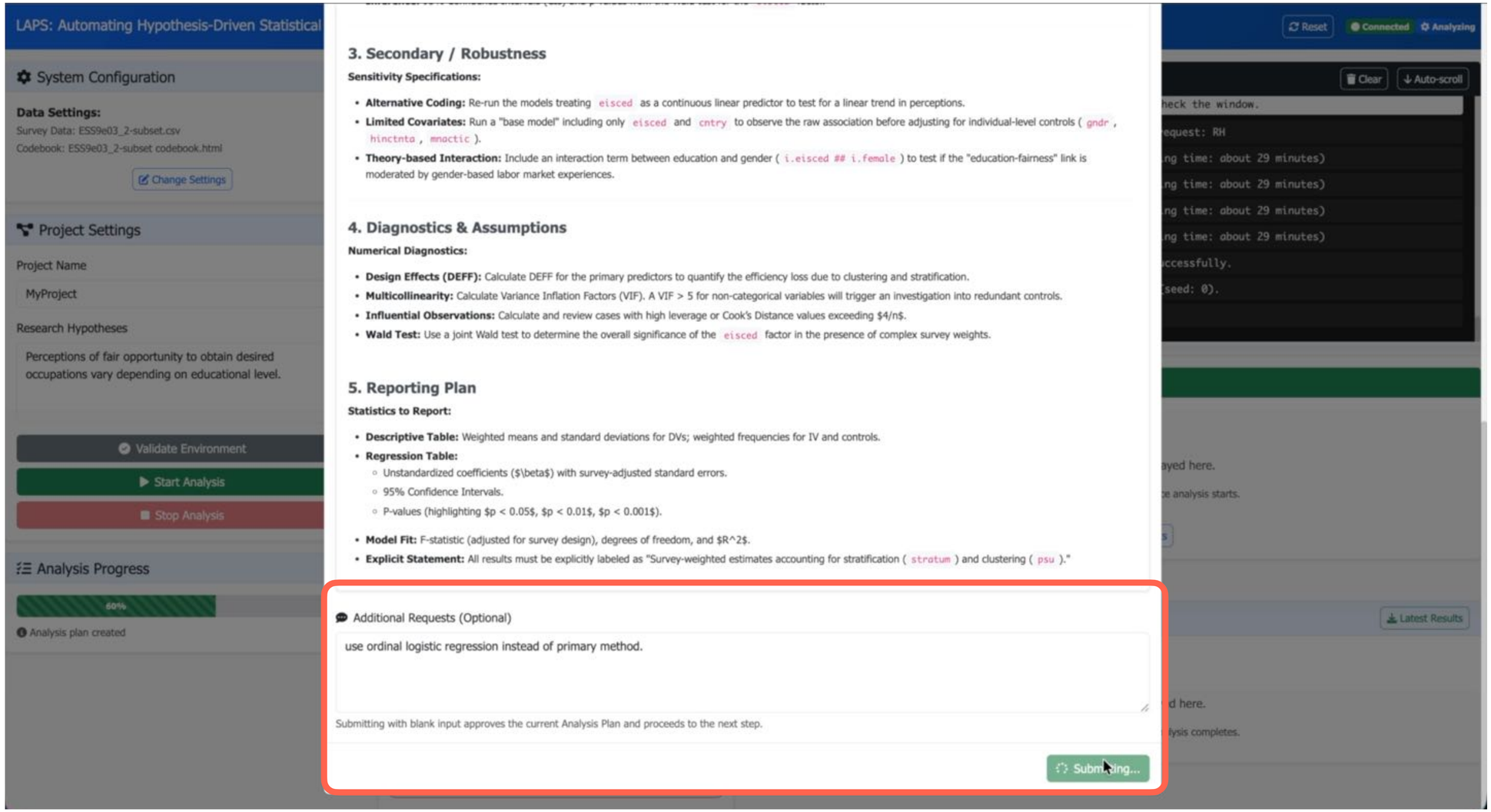
Step 1: Operationalization



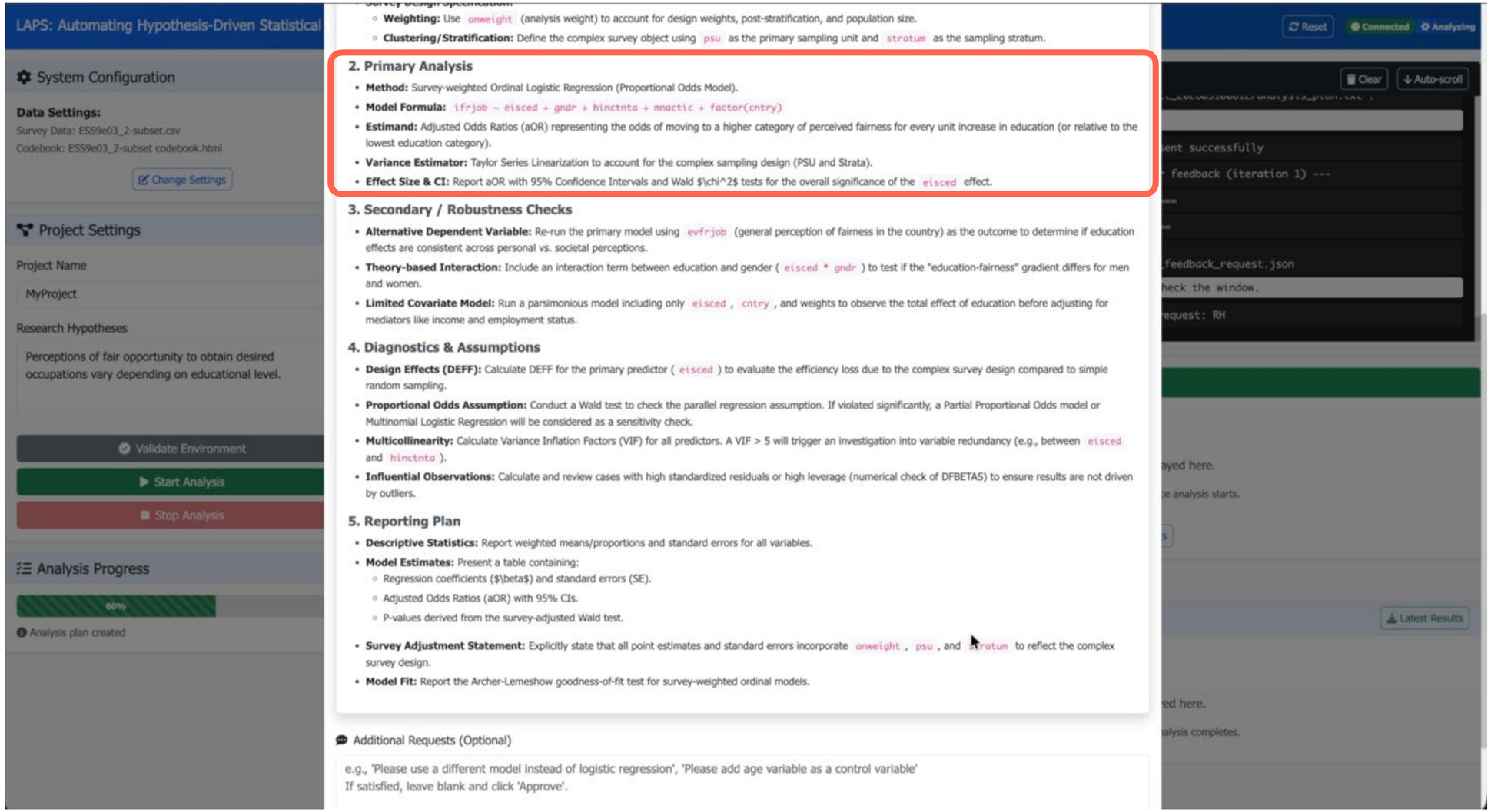
Step 1: Operationalization



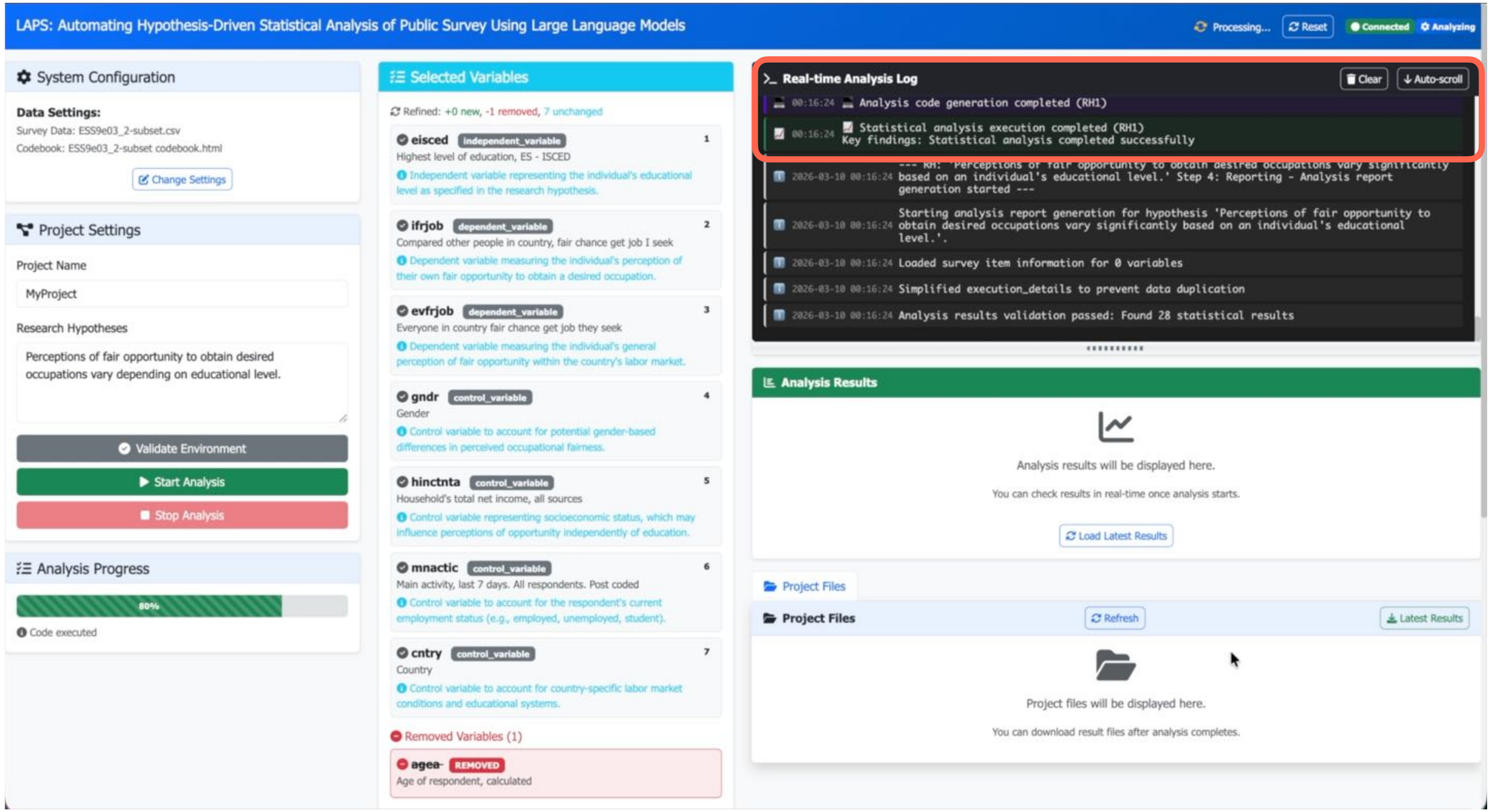
Step 2: Planning



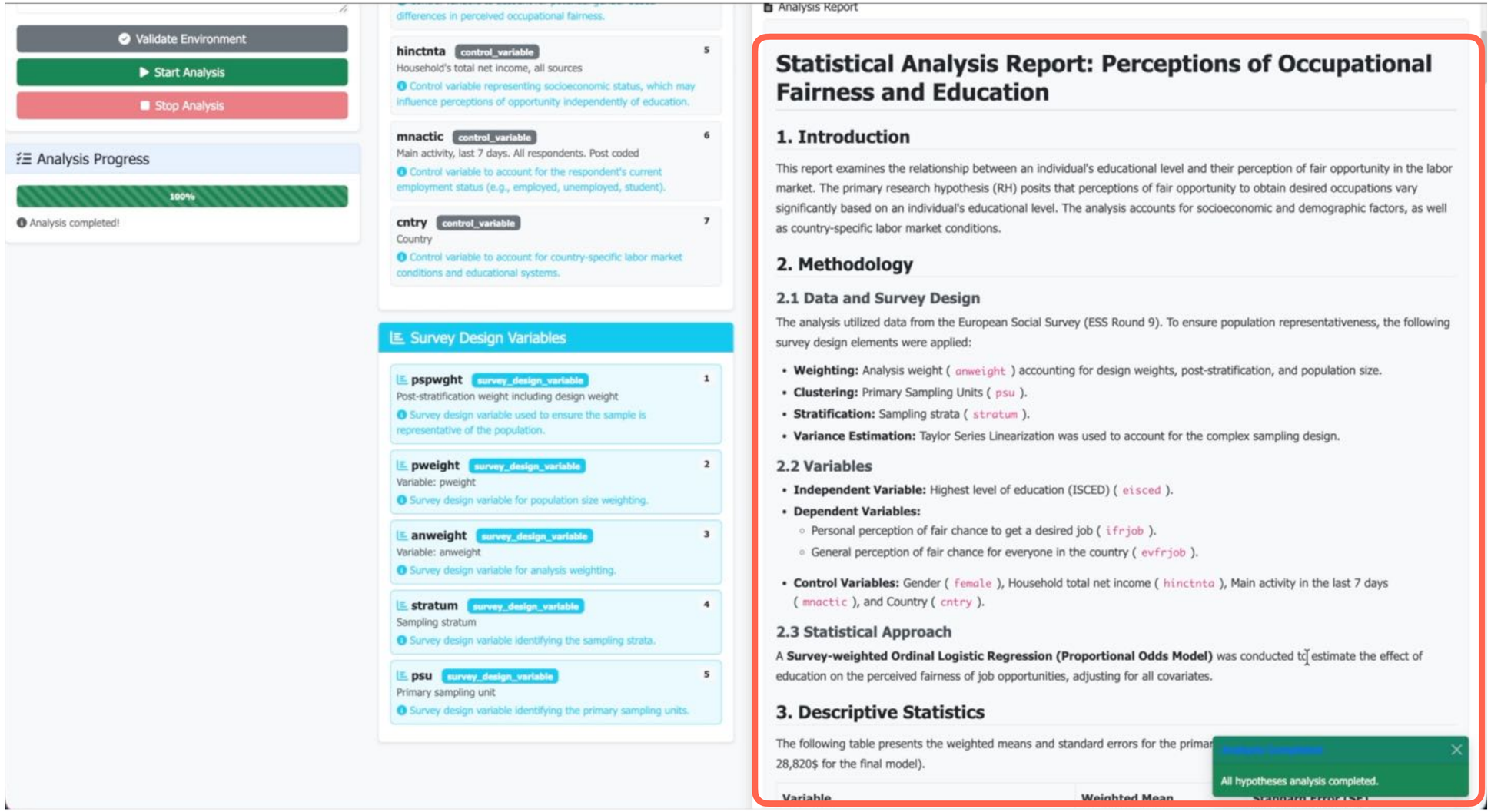
Step 2: Planning



Step 3: Execution & Step 4: Reporting



Step 3: Execution & Step 4: Reporting



Step 3: Execution & Step 4: Reporting

4. Primary Analysis Results

The survey-weighted ordinal logistic regression was performed with `ifrjob` (Personal fair chance) as the dependent variable.

4.1 Effect of Education

The analysis revealed a statistically significant positive relationship between education level and the perception of fair opportunity.

- **Education (`eisced`):** $\beta = 0.2044$, $SE = 0.0129$, Adjusted Odds Ratio (**aOR**) = **1.2268** (95% CI [1.1963, 1.2582]), $p < .001$.
- **Interpretation:** For every unit increase in the ISCED education level, the odds of reporting a higher level of perceived fair opportunity to obtain a desired job increase by approximately 22.7%, holding other factors constant.

4.2 Effects of Control Variables

All primary control variables showed significant associations with the dependent variable:

- **Gender (`female`):** $\beta = -0.1682$, aOR = 0.8452 (95% CI [0.7838, 0.9113]), $p < .001$. Females have significantly lower odds (approx. 15.5% lower) of perceiving a fair chance compared to males.
- **Income (`hinctnta`):** $\beta = 0.1058$, aOR = 1.1116 (95% CI [1.0935, 1.1299]), $p < .001$. Higher household income is associated with higher perceived fairness.
- **Main Activity (`mnactic`):** Compared to the reference category (Fact 1):
  - `mnactic_fact2` : aOR = 1.8665, $p < .001$ (Higher perceived fairness).
  - `mnactic_fact5` : aOR = 0.2793, $p < .001$ (Significantly lower perceived fairness).
  - `mnactic_fact6` : aOR = 0.6971, $p < .001$.
  - `mnactic_fact8` : aOR = 0.6349, $p < .001$.
  - `mnactic_fact7` and `fact9` were not statistically significant ($p = .3441$ and $p = .9896$, respectively).

4.3 Country Fixed Effects

Significant variation was observed across countries compared to the reference country:

- **Highest Perceived Fairness:** Switzerland (`cntry_factCH` , aOR = 2.0279) and Finland (`cntry_factFI` , aOR = 1.9046).
- **Lowest Perceived Fairness:** Montenegro (`cntry_factME` , aOR = 0.3065), Bulgaria (`cntry_factBG` , aOR = 0.3637), and Italy (`cntry_factIT` , aOR = 0.4148).

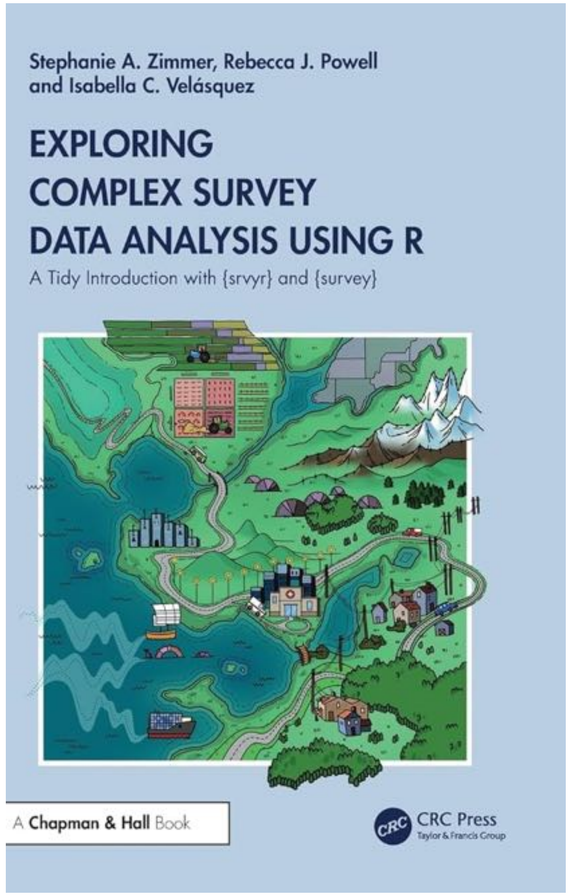
5. Diagnostics

- **Sample Size:** 28,820 observations were included in the primary model after cleaning and listwise deletion of missing values.
- **Design Effect:** The Design Effect (DEFF) for the primary predictor (`eisced`) was calculated to account for the complex survey design.
- **Model Fit:** The analysis was conducted using the `svyolr` function from the R `survey` package, ensuring that standard errors and p-values are adjusted for the multi-stage sampling design.

6. Conclusion

Technical Soundness

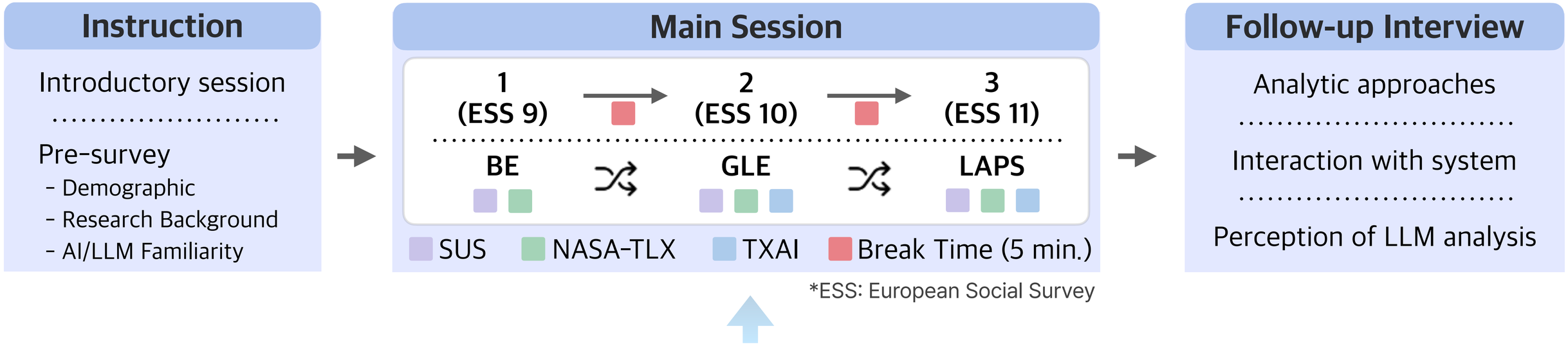
Evaluating five hypothesis-testing exercises from the Complex Survey Data Analysis textbook



Tutorial	Selected Variable	Statistical Method	Test Results
T1: Do more than 50% of U.S. households use air conditioning equipment?			
Answer	ACUsed	One-sample t-test	$p < 0.001$
LAPS	ACUsed	One-sample z-test	$p < 0.001$
T2: Does the average temperature that U.S. households set their thermostats to differ between the day and night in the winter?			
Answer	WinterTempDay; WinterTempNight	Paired t-test	$p < 0.001$
LAPS	WinterTempDay; WinterTempNight	Paired t-test	$p < 0.001$
T3: Is there a relationship between the type of housing unit and the year the house was built?			
Answer	HousingUnitType; YearMade	Wald chi-square test	$p < 0.001$
LAPS	HousingUnitType; YearMade	Pearson chi-square test	$p < 0.001$
T4: Does the average age of those who voted for Joseph Biden in 2020 differ from those who voted for another candidate?			
Answer	(DV) Age; (IV) VotedPres2020_selection	Two-sample t-test	$p < 0.001$
LAPS	(DV) Age; (IV) VotedPres2020_selection	Two-sample t-test	$p < 0.001$
T5: Is there a difference in the distribution of gender across early voting status in 2020?			
Answer	Gender, EarlyVote2020	Rao–Scott chi-square test	$p = 0.03$
LAPS	Gender, EarlyVote2020	Rao–Scott chi-square test	$p = 0.03$

Study Procedure

Formulating hypotheses and conducting a 4-step analysis workflow



“Within-Subjects Design”

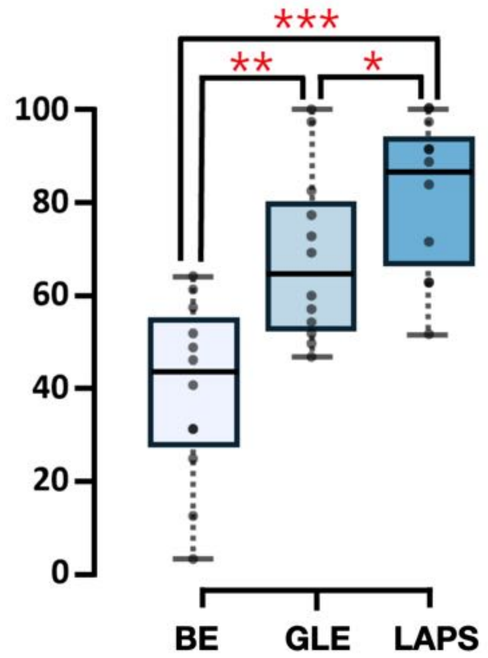
1) Baseline (BE): Traditional tools only (SPSS, R) + Web Search
2) General LLM (GLE): Traditional tools + ChatGPT 5.1
3) LAPS: Our proposed framework

Participants

ID	Gender	Age	Education	Research Experience (Years)	Round 1 (ESS 9)	Round 2 (ESS 10)	Round 3 (ESS 11)
P1	Male	31	Master's Student	3	BE	GLE	LAPS
P2	Female	33	PhD Candidate	8	BE	GLE	LAPS
P3	Male	53	PhD	25	BE	LAPS	GLE
P4	Female	32	PhD Student	4	BE	LAPS	GLE
P5	Male	29	PhD Student	4	GLE	BE	LAPS
P6	Female	25	PhD Student	2	GLE	BE	LAPS
P7	Female	40	PhD Student	6	GLE	LAPS	BE
P8	Female	33	PhD Candidate	4	GLE	LAPS	BE
P9	Female	51	PhD	20	LAPS	BE	GLE
P10	Female	27	PhD Student	3	LAPS	BE	GLE
P11	Female	33	PhD Candidate	12	LAPS	GLE	BE
P12	Female	31	PhD Candidate	6	LAPS	GLE	BE

Ensuring Researcher Agency and Analytical Stability

System Usability Scale



LAPS as a collaborative tool that preserves the researcher's control over key analytical decisions

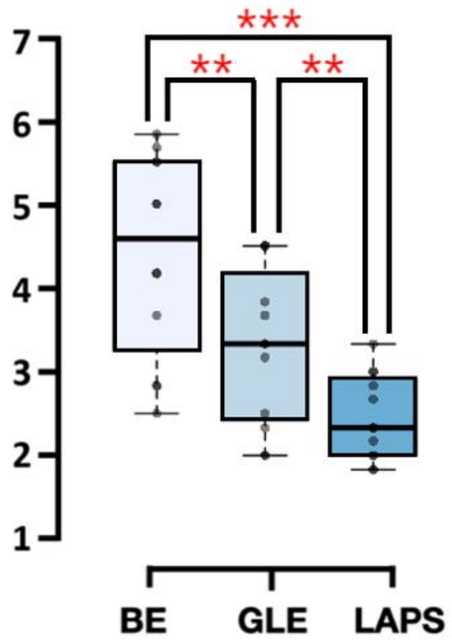
"I could revise and adjust variable selection and the analysis plan [...] and it always felt like the tool left the decision to me. I could use the suggested method as it was or combine multiple options, so it really felt like it was giving me choices" (P3)

LAPS raises the baseline of analysis quality

"The biggest barrier in social science is the time it takes to learn statistical tools [...] and a tool like LAPS would let students or researchers try more kinds of studies and make research more accessible" (P7)

Reducing Cognitive Burden in the Analysis Workflow

NASA-TLX



LAPS reduces the cost of exploratory analysis by making early analytical conditions visible

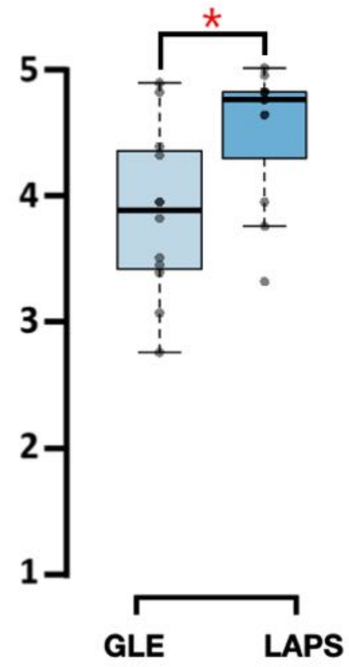
"Social science research requires constant movement between theory and exploratory analysis. [...] LAPS helps reduce the cost of testing analytic ideas quickly, allowing me to check whether an analysis is feasible with the dataset before committing to a new method" (P11)

LAPS enables researchers to focus on conceptual decisions rather than tool management

"LAPS automatically organizes everything I used to stitch together on my own, collecting the variables, setting up the models, and summarizing the logic behind each step. [...] It saved time, but more importantly, it freed up mental bandwidth because I didn't have to manage the process myself constantly" (P5)

Trust in LAPS Results

TXAI



LAPS builds confidence by providing transparent, domain-bounded reasoning rather than black-box recommendations

"With general LLMs, I often felt anxious about hallucinated variables or arbitrary statistical choices. LAPS, however, clearly grounds its recommendations in the actual survey metadata, which made me trust its output" (P2)

"Because the system explicitly explains why it chose specific variables and statistical methods, I didn't feel like I was blindly accepting an AI's answer. It felt like a transparent collaboration" (P10)

Design Implications for Expert Workflows

Supporting Early-Stage Researchers

LAPS mitigates analysis quality variation, but novices may still overlook essential data verification steps

↓

Reconceptualize data processing as visible, traceable decisions and provide reflective cues to maintain data literacy

Integrating Knowledge Resources

LAPS sustains a coherent workflow, but researchers occasionally need an external theoretical context

↓

Integrate “on-demand evidence” features to link choices with literature without causing information overload

Interpretive Variability

AI explanations are helpful, but meaningful interpretations inherently vary across disciplinary traditions

↓

Introduce a cross-domain simulation agent to present diverse perspectives while preserving researcher agency

Broader Impacts & Procedural Responsibility

The Double Edge of Safety

Strict LLM safety filters often block legitimate analyses of sensitive public survey items

↓

Balance model selection and safety policies to ensure analytic robustness for sensitive research topics

Procedural Responsibility

Research reliability and procedural transparency fundamentally depend on the researcher's ethical commitment

↓

Automate audit trails and workflow logs to facilitate transparent, evidence-based reporting

CHI'26

HANYANG UNIVERSITY

CHUNGBUK NATIONAL UNIVERSITY

LAPS:
Automating Hypothesis-Driven Statistical Analysis of Public Survey Using Large Language Models

PAPER

Thank You


Jaehoon Kim*


Dayoung Jeong*


Beejin Son


Hansung Kim


Bogoan Kim


Kyungsik Han

** Both authors contributed equally to this research*

Contact: {jaehoonkimm, dayoungjeong, kyungsikhan} @ hanyang.ac.kr

HANYANG HUMAN CENTERED COMPUTING LAB

Appendix - Step 1

Codebook Parser

You are an expert in survey codebook analysis with deep knowledge of various codebook formats. Your task is to extract comprehensive variable information from this codebook, including detailed survey questions and response choice labels.

 # INPUT:

`Codebook Data: {codebook_text}`

 # TASK INSTRUCTION

(1) Variable Names: Look for various patterns:
- Standard formats: V200001, V201006, Q123
- Descriptive names: Age, Gender, Education, Income
- Mixed formats: CaseID, InterviewMode, Weight

(2) Variable Descriptions/Labels: Extract detailed descriptions:
- Full descriptive text found near variable names
- Category labels and value descriptions
- Scale descriptions (1-5, 1-7, etc.)
- Multiple choice option labels

(3) Survey Question Text WITH Response Options: Extract complete question wording:
- Full question text as asked to respondents
- Include ALL response choice labels in the question text

(4) Response Choice Labels Integration: always include when available:
- All response option labels (e.g., "1. Strongly disagree", "2. Disagree")
- Value-label pairs for categorical variables (e.g., "1=Video, 2=Telephone, 3=Web")
- "Don't know", "Refused", "Not applicable" options

 # OUTPUT

Return results as a JSON array. For each variable found, include response options in the survey_question:

```
[
    {
        "variable_name": "V200002",
        "indicator": "Mode of interview for pre-election survey - How the respondent participated in
      the survey",
        "survey_question": "How did you participate in this pre-election survey interview? [Response
      options: 1=Video, 2=Telephone, 3=Web]"
    }, ...
]
```

Hypothesis Clarifier

You are an expert specializing in hypothesis clarification. Your task is to:
1. Check if the user's input is a valid research hypothesis.
2. If valid, identify a research hypothesis (RH) from the user input and return it as a JSON array. The hypothesis should have a "hypothesis" key.
3. If invalid, return "not_hypothesis" with an "error" key and explain why.

 # INPUT:

`User Input: {user_input}`
`Dataset Information: {dataset_information}`

 # INVALID INPUT EXAMPLE

- Questions or casual conversation
- Unclear/meaningless text, random words, incomplete sentences
- Non-research statements (opinions, facts, descriptions without testable predictions)
- Hypotheses that cannot be analyzed using the dataset, based on the "Dataset Information"

 # TASK INSTRUCTION

- Analyze the input and return a JSON array.
- If the input is a VALID research hypothesis (or can be reasonably interpreted as one):
  - User input: "Adolescence is a period of significant physical and psychological change, and various problems can arise as a result. This study aims to determine the potential impact of adolescents' eating habits on their mental health."
  - Output: ["hypothesis": "Adolescent dietary habits will affect adolescent mental health."]
- If the input is INVALID (not a research hypothesis):
  - User input: "What is mental health?"
  - Output: ["error": "not_hypothesis", "message": "This is a general question, not a testable research hypothesis."]

Appendix - Step 1

Variable Selector

You are a statistical analysis expert.
If this is the first attempt:

Please select appropriate variables suitable for research hypothesis analysis according to the given instructions.

If this is the refinement:

Based on the feedback provided, please refine the variable selection and return an improved variable list.

 # INPUT:

```
User Input: {user_input}
Variable List: {survey_item_list}
```

 # TASK INSTRUCTION

- Variable Selection Guideline
  - Select ONLY variables that exist in the provided survey item list.
  - NEVER create or invent variable names that are not present in the survey data.
  - If multi-item variables are selected, the entire set must be included.
  - variable_role must be assigned to one of 'independent_variable', 'dependent_variable', 'control_variable', or 'survey_design_variable'.
- VARIABLE DUPLICATION PREVENTION
  - NEVER use the same variable for multiple roles (e.g., as both dependent and control variable)
  - Each variable should have only ONE role assignment
  - If a variable could serve multiple purposes, choose its PRIMARY role based on the hypothesis
  - If 'depression' could be dependent variable OR control variable, choose based on hypothesis focus
  - Check your final list - no variable_name should appear more than once

 # EXAMPLE

```
[
    {{
        "indicator": "Gender",
        "variable_name": "SEX",
        "survey_question": "Gender",
        "explanation": "Independent variable measuring 'gender' mentioned in the hypothesis",
        "variable_role": "independent_variable"
    }}, ...
]
```

Variable Completer

You are a statistical analysis expert. You will validate the variables required for the selected hypothesis test and provide focused feedback.

 # INPUT:

```
Research Hypothesis: {hypothesis}
All Variable List: {survey_questions_str}
Currently Selected Variable List: {selected_variables}
```

 # TASK INSTRUCTION

When reviewing the variable selection, provide comprehensive feedback by checking:

(1) Missing or Required Variables
- Compare the All Variables list with the Currently Selected Variables list.
- Identify any variables required for hypothesis testing that are missing.

(2) Unnecessary Variables
- Check if the Currently Selected Variables list contains any variables irrelevant to the hypothesis.
- Survey design variables must be maintained.

(3) Selecting a non-existent variable
- Check if the Currently Selected Variables list contains a non-existent variable in the All variable list.
- If a non-existent variable is selected, feedback must be provided to correct it to the variable's valid name (as listed in the All Variable List).

(4) Multi-item Scales (e.g., GAD_1–GAD_7, PHQ_1–PHQ_9, SES_1–SES_5)
- If one item from a scale is selected, the entire set must be included.
- All items with the same prefix are included.
- The number of selected items matches the known set size.

Ensure the feedback is clear, comprehensive, and points out both omissions and unnecessary inclusions. Output must be in JSON array format.

 # OUTPUT

```
[
    {{
        "Feedback": "YOUR FEEDBACK"
    }}
]
```

Appendix - Step 2

Data Descriptor

You are a statistical analysis expert. Please interpret the descriptive statistics results for the following research hypothesis and variables.

 # INPUT:

```
Research Hypothesis: {hypothesis_text}
Selected Variable Information: {variables_summary}
Descriptive Statistics Results: {stats_summary}
```

 # TASK INSTRUCTION

- Distribution Characteristics of Each Variable: Briefly describe the distribution shape, central tendency, and dispersion of each variable.
- Observed Group Differences: Explain how the distribution or proportions of dependent variables differ across categories of independent variables.
- Preliminary Hypothesis-Related Observations: Describe patterns or trends related to the research hypothesis based solely on descriptive statistics.
- Data Quality Assessment: Mention key features related to data quality such as missing values, outliers, and distributional skewness.
- Important Notes
  - Do not mention statistical significance (as testing has not been performed yet).
  - Only provide descriptive commentary on observed patterns or differences.
  - Use specific numerical values in your explanations.

Please write in markdown format.

Analysis Planner

You are a statistical expert in complex survey analysis. Draft a focused, executable statistical analysis plan to test the RH using appropriate survey methods.

 # INPUT:

```
Research Hypothesis: "{hypothesis}"
Selected Variables: {variables_str}
Descriptive Statistics: {descriptive_stats_str}
Survey Design Variables: {survey_design_str}
```

 # TASK INSTRUCTION

- Consider complex survey design (weights/PSU/strata).
- Identify all variables requiring recoding with specific recoding steps and plan recoding if needed.
- Prioritize a direct test of the primary hypothesis.
- Select statistical models consistent with available sample size and missingness.
- Important Notes
  - Use exact variable names from “variables_str”.
  - If any selected variable is excluded or its role differs, briefly justify.
  - Keep the plan clear and directly executable.

 # OUTPUT

(1) Data pre-processing & coding: Recodes (with rules), inclusion/exclusion, missing-data approach.
(2) Primary analysis: Method & model (survey-aware), formula with exact variable names, estimand, variance estimator, and effect size & CI.
(3) Secondary / robustness: Key sensitivity specs (e.g., alternative codings, limited covariates, stratified checks) and any single, theory-based interaction.
(4) Diagnostics & assumptions: Design effects, fit/linearity where relevant, and influential observations.
(5) Reporting plan: Statistics to report (estimates, CIs, p-values, effect sizes), with survey adjustment explicitly stated.

Appendix - Step 3

Code Generator

You are an expert statistical programming engineer specialized in end-to-end survey analysis using Python and R. Generate Python code that uses rpy2 to execute a complete R analysis script as a single block.

 # INPUT:

```
Analysis Plan: {analysis_plan}
Data Path: {survey_data_path}
Selected Variables: {selected_variables}
Survey Design Variables: {survey_design_variables}
```

 # TASK INSTRUCTION

(1) Write a complete R script as a string that performs the entire analysis
(2) Use rpy2's ro.r() to execute the entire R script at once
(3) The R script should handle all data loading, processing, analysis, and result formatting
(4) Minimize Python-R object conversions - do everything in R and return final JSON results
(5) Use the exact variable names provided in Selected Variables

Generate the complete Python code with an embedded R script that performs the entire statistical analysis described in the plan. The R script should be self-contained and handle all data type conversions, survey design setup, and analysis within R.

Code Debugger

You are an expert in code debugging, specializing in Python-R code integration using the rpy2 library. Analyze the error and write the feedback for debugging.

 # INPUT:

```
Available libraries: {available_libs}
Data samples: {data_sample_context}
Refinement history: {refinement_history_context}
Selected Variables: {variables_context}
Current Code: {code_context_str}
Analysis Plan: {analysis_plan}
Execution Error Information: {full_error_context}
```

 # TASK INSTRUCTION

Analyze the error and write the feedback for debugging.

(1) Review the data samples above to understand data formats, encodings, and column names that may be causing issues.
(2) Check previous debugging attempts to avoid repeating failed approaches - suggest different solutions if similar fixes were tried before.
(3) If validation results show this is a recurring error pattern, provide a completely different approach rather than incremental fixes.
(4) Determine error type and provide specific, actionable feedback with exact code changes needed.

 # OUTPUT

Provide detailed, actionable debugging feedback:

- Root Cause: [Brief description of the root cause with reference to specific line/variable.]
- Problem: [What exactly is wrong]
- Solution: [Exact code changes needed with line numbers. If this is a recurring issue, provide an alternative implementation approach.]

Appendix - Step 4

Report Composer

You are a statistical analysis expert. Generate a statistical analysis report based on the provided analysis results.

 # INPUT:

```
Research Hypothesis: {hypothesis_text}
Selected Variables: {selected_variables}
Survey Items and Questionnaire Information: {survey_items}
Analysis Plan: {analysis_plan}
Statistical Analysis Results: {analysis_results}
Descriptive Statistics: {descriptive_stats}
```

 # TASK INSTRUCTION

(1) NEVER create, invent, or guess ANY statistical values (p-values, test statistics, coefficients, etc.)
(2) ONLY use numbers and results that are EXPLICITLY present in the Statistical Analysis Results above
(3) If a statistical test result is missing or shows an error, state "Result not available due to analysis error"
(4) Do NOT fill in missing results with placeholder values or estimates
(5) Use exact quotes when reporting error messages from failed analyses
(6) If multiple variables were analyzed, report ALL of them systematically, not just selected examples
(7) Use the "Survey Items and Questionnaire Information" above to correctly interpret variable meanings and direction of effects
- For each variable in your analysis, refer to the survey question to understand
  - What the question actually asked respondents
  - The response scale and direction (e.g., 1=strongly disagree to 5=strongly agree)
  - Whether higher values indicate more or less of the construct
- Do not assume the direction of effects without checking the actual survey questions - When interpreting coefficients, explicitly reference the survey question to explain the direction

 # REPORTING STANDARDS (APA Guidelines)

Follow APA guidelines for reporting ONLY the statistics that are available.

(1) ALWAYS include if descriptive statistics are provided - report sample sizes, means, proportions, standard deviations for ALL variables
(2) Report EVERY analysis conducted, not just selected results. For each analysis result, report:
- The statistical method used
- The exact test statistic value
- The exact p-value
- Effect size ONLY if calculated and present in results
- Confidence intervals ONLY if present in results

(3) If an analysis failed or has errors, explicitly state this in the report
(4) Clearly distinguish between successful and failed analyses
(5) Do not use phrases like "For example" or "Among the significant results" - report ALL results systematically
(6) Include both significant AND non-significant results in your analysis - both are scientifically important

Generate a professional statistical analysis report based on these results and requirements in markdown format.