**Assessing Coding Proficiency and Educational Background of Graduate Students with Beginner to Intermediate Comfort in Coding Enrolled in the Advanced Data Analysis Course at the Brown School**

**Methods**

**Study Design:** This was a cross-sectional study.

**Study Setting:** Data were collected remotely via a Google Forms survey accessible to anyone with the link from August 12th to 27th, 2024.

**Eligibility Criteria:** Participants were students enrolled in the Advanced Data Analysis Fall 2024 class at Washington University in St. Louis, Brown School. The survey link was accessible via the class Canvas page and emailed to all students.

**Variables:** Participants' chronotype was determined by asking, "Are you a lark, an owl, or a hummingbird?" The options were: Lark (I am a morning person), Owl (I am a night person), and Hummingbird (I am neither a lark nor an owl). For analysis purposes, we classified those who selected "Owl" as the owl group and combined "Lark" and "Hummingbird" responses into a non-owl group.

The coding comfort level was assessed with the question, "How comfortable are you with coding?" The scale ranged from Not comfortable (1) to Very comfortable (5). Participants who rated themselves from 1 to 3 were considered to have beginner to intermediate coding experience for our analysis, while those who rated 4 or 5 were considered advanced. Our focus was on participants with coding comfort levels of 3 or less, aligning with our interest in beginner to intermediate coders.

Additional variables included birth month, collected by asking "What is your birth month?" with numeric responses corresponding to months (e.g., January = 1). We also gathered information on current degree program (e.g., MPH, MSW), highest level of education (< High school, High school, bachelor's, master's, doctoral), primary statistical software (SPSS, Excel, R, Stata), level of R experience (using the same 1 to 5 scale as coding comfort), and duration of software use (options like "0–3 months," "4–6 months"). These variables were measured directly from survey responses and analyzed to understand participants' educational backgrounds and software proficiency levels.

groups.

**Bias:** Data collection via a self-administered online survey may introduce recall bias due to inaccurate self-reporting, such as overestimating coding proficiency. We did not adjust for this bias. Stratified subgroup analyses between Global Health and Epidemiology/Biostatistics concentrations among owls and non-owls helped control for confounding variables. Residual bias may remain due to the absence of sensitivity analyses.

**Study Size:** From 29 survey respondents, we identified 8 owls and 11 non-owls with coding comfort levels of 3 or below after applying exclusion criteria. We excluded participants not matching the owl/non-owl categories, those with missing or non-numeric birth month data, and those with coding experience over three years to maintain consistent skill levels. The online survey consisted of 27 questions.

**Statistical Analysis:** We used R software (version 4.4.1) for all statistical analyses. Descriptive statistics were calculated to compare owls and non-owls across key variables: degree program, software preference, level of R experience, duration of software use, and highest level of education. To control for potential confounders, we performed stratified analyses based on specialization areas (Global Health and Epidemiology/Biostatistics). Assumptions for statistical tests were assessed to ensure validity. No adjustments for missing data were necessary, as there were no missing values in our variables of interest. Sensitivity analyses were not conducted due to the small sample size.

**Code Availability Statement:** The source code used for this assignment can be obtained using this R file: <https://github.com/dtsiamtsiouris/HW4>

**Results**

We analyzed data from 19 participants with coding comfort levels of 3 or less, comprising 8 owls and 11 non-owls. The majority were enrolled in the MPH program (89.5%), with all non-owls (100%) and most owls (75%) in this program. Most participants held a bachelor's degree (68.4%). SPSS was the most preferred software overall (42.1%), favored by 45.5% of non-owls and 37.5% of owls, followed by Excel (21.1%), R, and Stata (each 15.8%). Owls displayed a varied distribution in the duration of software use, while non-owls predominantly reported 4–6 months of use (36.4%).

Owls reported slightly higher coding comfort levels in R (mean = 2.63, SD = 0.74) compared to non-owls (mean = 2.27, SD = 0.65), with medians of 3.00 and 2.00 respectively. This suggests owls felt more comfortable with coding within the beginner to intermediate range. Stratified analyses by specialization (Global Health vs. Epidemiology/Biostatistics) revealed minimal differences between owls and non-owls in software preferences and coding comfort levels within these subgroups.

**Characteristics of the Study population:** The study population of 19 participants was divided into owls and non-owls based on self-reported active times, enabling comparative analysis. We examined variables like degree program, level of education, software preferences, level of R experience, and duration of software use to understand participants' educational backgrounds and software skills.

[Please see Appendix for detailed flowchart showing the derivation of our analytic dataset for the owls & non-owls groups.]

**Table 1 Summary:** This table highlights the key attributes of the two groups. Both groups have similar educational backgrounds, with most holding an MPH degree only. In terms of software usage, Owls are more likely to use R (25%) and Stata (25%), while Non-Owls are more inclined toward Excel (27.3%) and SPSS (45.5%). The Rx (level of R experience) mean score is slightly higher for Owls (2.63) compared to Non-Owls (2.27), indicating greater R experience in the Owl group. When considering the duration of coding, 25% of Owls have coding experience between 0-9 months, falling into multiple categories, unlike Non-Owls, who are more concentrated in the 4-6 months and 9 months to 1 year ranges. Regarding the level of education (LOE), the majority of both Owls (75%) and Non-Owls (63.6%) hold a bachelor's degree, with smaller proportions holding master's or doctoral degrees in both groups.

[Please see Appendix for Table 1]

**Table 2 Summary:** This table compares Owls and Non-Owls by their specialization in Epidemiology/Biostatistics and Global Health. Most participants in Epidemiology/Biostatistics hold an MPH degree only, while in Global Health, half of the Owls hold an MPH/MSP. Non-Owls in Epidemiology/Biostatistics favor SPSS (45.5%) and Excel (27.3%), while Owls use more R (33.3%) and SPSS (33.3%). In Global Health, Owls use SPSS and Stata equally (50%). The Rx (R experience) scores are higher for Global Health Owls (mean: 3.00) compared to Epidemiology/Biostatistics (mean: 2.50 for Owls and 2.27 for Non-Owls). Owls in Epidemiology/Biostatistics show more varied short-term coding experience (0-9 months), while Non-Owls are spread across longer durations. Regarding education, the majority hold bachelor’s degrees, but Global Health Owls have higher levels of education, with 50% holding doctoral degrees.

[Please see Appendix for Table 2]

**Appendix**

**Fig 1:** Flowchart showing the derivation of our analytic dataset for the owls & non-owls groups

A screenshot of a diagram

Description automatically generated

**Table 1: Summary Characteristics of Owls and Non-Owls**

|  | **Owl (N=8)** | **Non-owl (N=11)** | **Overall (N=19)** |
| --- | --- | --- | --- |
| **degree** |  |  |  |
| MPH only | 6 (75.0%) | 11 (100%) | 17 (89.5%) |
| MPH/MSP | 1 (12.5%) | 0 (0%) | 1 (5.3%) |
| Missing | 1 (12.5%) | 0 (0%) | 1 (5.3%) |
| **software** |  |  |  |
| Excel | 1 (12.5%) | 3 (27.3%) | 4 (21.1%) |
| R | 2 (25.0%) | 1 (9.1%) | 3 (15.8%) |
| SPSS | 3 (37.5%) | 5 (45.5%) | 8 (42.1%) |
| Stata | 2 (25.0%) | 1 (9.1%) | 3 (15.8%) |
| I've used both R and SPSS | 0 (0%) | 1 (9.1%) | 1 (5.3%) |
| **Rx** |  |  |  |
| Mean (SD) | 2.63 (0.744) | 2.27 (0.647) | 2.42 (0.692) |
| Median [Min, Max] | 3.00 [1.00, 3.00] | 2.00 [1.00, 3.00] | 3.00 [1.00, 3.00] |
| **duration** |  |  |  |
| >1 to 2 years | 1 (12.5%) | 1 (9.1%) | 2 (10.5%) |
| 0-3 months | 2 (25.0%) | 1 (9.1%) | 3 (15.8%) |
| 4-6 months | 2 (25.0%) | 4 (36.4%) | 6 (31.6%) |
| 7-9 months | 2 (25.0%) | 2 (18.2%) | 4 (21.1%) |
| 9 months to 1 year | 1 (12.5%) | 3 (27.3%) | 4 (21.1%) |
| **LOE** |  |  |  |
| Bachelor's degree | 6 (75.0%) | 7 (63.6%) | 13 (68.4%) |
| Doctoral degree | 1 (12.5%) | 2 (18.2%) | 3 (15.8%) |
| Master's degree | 1 (12.5%) | 2 (18.2%) | 3 (15.8%) |

**Table 2: Summary Characteristics of Owls and Non-Owls by Their Specialization**

|  | **Epidemiology/Biostatistics** | | | **Global Health** | **Overall** | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Non-Owls (N=11)** | **Owls (N=6)** | **Owls (N=2)** | | **Non-Owls (N=11)** | **Owls (N=8)** |
| **degree** |  |  |  | |  |  |
| MPH only | 11 (100%) | 6 (100%) | 0 (0%) | | 11 (100%) | 6 (75.0%) |
| MPH/MSP | 0 (0%) | 0 (0%) | 1 (50.0%) | | 0 (0%) | 1 (12.5%) |
| Missing | 0 (0%) | 0 (0%) | 1 (50.0%) | | 0 (0%) | 1 (12.5%) |
| **software** |  |  |  | |  |  |
| Excel | 3 (27.3%) | 1 (16.7%) | 0 (0%) | | 3 (27.3%) | 1 (12.5%) |
| I've used both R and SPSS | 1 (9.1%) | 0 (0%) | 0 (0%) | | 1 (9.1%) | 0 (0%) |
| R | 1 (9.1%) | 2 (33.3%) | 0 (0%) | | 1 (9.1%) | 2 (25.0%) |
| SPSS | 5 (45.5%) | 2 (33.3%) | 1 (50.0%) | | 5 (45.5%) | 3 (37.5%) |
| Stata | 1 (9.1%) | 1 (16.7%) | 1 (50.0%) | | 1 (9.1%) | 2 (25.0%) |
| **Rx** |  |  |  | |  |  |
| Mean (SD) | 2.27 (0.647) | 2.50 (0.837) | 3.00 (0) | | 2.27 (0.647) | 2.63 (0.744) |
| Median [Min, Max] | 2.00 [1.00, 3.00] | 3.00 [1.00, 3.00] | 3.00 [3.00, 3.00] | | 2.00 [1.00, 3.00] | 3.00 [1.00, 3.00] |
| **duration** |  |  |  | |  |  |
| >1 to 2 years | 1 (9.1%) | 0 (0%) | 1 (50.0%) | | 1 (9.1%) | 1 (12.5%) |
| 0-3 months | 1 (9.1%) | 1 (16.7%) | 1 (50.0%) | | 1 (9.1%) | 2 (25.0%) |
| 4-6 months | 4 (36.4%) | 2 (33.3%) | 0 (0%) | | 4 (36.4%) | 2 (25.0%) |
| 7-9 months | 2 (18.2%) | 2 (33.3%) | 0 (0%) | | 2 (18.2%) | 2 (25.0%) |
| 9 months to 1 year | 3 (27.3%) | 1 (16.7%) | 0 (0%) | | 3 (27.3%) | 1 (12.5%) |
| **LOE** |  |  |  | |  |  |
| Bachelor's degree | 7 (63.6%) | 5 (83.3%) | 1 (50.0%) | | 7 (63.6%) | 6 (75.0%) |
| Doctoral degree | 2 (18.2%) | 0 (0%) | 1 (50.0%) | | 2 (18.2%) | 1 (12.5%) |
| Master's degree | 2 (18.2%) | 1 (16.7%) | 0 (0%) | | 2 (18.2%) | 1 (12.5%) |

SD = Standard Deviation