# Understanding Student Interest in Substance Use Disorder Counseling: A Mixed-Methods Approach Using Machine Learning and Qualitative Interviews

Erika Hernandez Acton, Linda Reynolds, Bryan Acton, Rory Eckhardt, Giovanni Scaringi, and Nadine Mastroleo

Binghamton University

# Author Note

Correspondence concerning this article should be addressed to Erika Hernandez Acton, Binghamton University, Email: [erika@binghamton.edu](mailto:erika@binghamton.edu)

# Abstract

*Keywords*: substance use disorder, counseling, workforce development, student interest, career choice, machine learning, random forest, qualitative interviews, mixed-methods

# Understanding Student Interest in Substance Use Disorder Counseling: A Mixed-Methods Approach Using Machine Learning and Qualitative Interviews

# Introduction

Substance use disorders (SUDs) continue to pose a major public health challenge, creating a critical need for skilled counselors to provide addiction treatment. The U.S. Bureau of Labor Statistics projects a 19% increase in employment of substance abuse and mental health counselors from 2023 to 2033—nearly five times the average job growth across all occupations ([U.S. Bureau of Labor Statistics, 2024](#ref-bls2024)). This translates to approximately 49,000 job openings each year for counselors in this sector. The opioid epidemic, combined with rising rates of alcohol use disorders and polysubstance use, has intensified demand for qualified professionals who can deliver evidence-based addiction treatment. Yet the pipeline of new professionals entering SUD counseling has not kept pace with this urgent demand, creating significant workforce shortages that limit access to care for millions of Americans struggling with addiction.

## Workforce Shortage and Recruitment Challenges

Despite substantial investment in workforce development—with federal agencies allocating millions annually to recruitment and training initiatives—the SUD counseling field continues to face critical staffing shortages. These shortages are particularly acute in rural areas and underserved communities, where the need for addiction services is often greatest ([Health Resources and Services Administration, 2023](#ref-hrsa2023)). The challenge extends beyond simple numbers; the field struggles with high turnover rates, with studies indicating annual turnover approaching 50% in some treatment settings ([Eby & Rothrauff-Laschober, 2010](#ref-eby2010)). This combination of recruitment difficulties and retention challenges creates a compounding effect that undermines the stability and effectiveness of the addiction treatment system.

Several factors contribute to these workforce challenges. SUD counseling has historically carried stigma within the broader mental health field, often viewed as less prestigious or professionally rewarding than other counseling specialties ([Knox et al., 2019](#ref-knox2019)). Educational programs have traditionally provided minimal addiction-specific training, with many counseling and psychology programs offering little more than a single elective course on substance use issues. This educational deficit has been identified as a “significant roadblock” to expanding the SUD workforce, as students receive limited exposure to the field during their formative training years ([Renner et al., 2019](#ref-renner2019)). Furthermore, misconceptions about the emotional toll of working with individuals with SUDs—including beliefs about inevitable relapse and treatment resistance—may deter students from considering this career path.

## Factors Influencing SUD Career Interest

Understanding what draws individuals to or deters them from SUD counseling careers is essential for developing effective recruitment strategies. Prior research suggests several key factors that influence career choice in this field. Personal or family experience with substance use disorders emerges consistently as a motivating factor, with studies indicating that a substantial proportion of addiction counselors have either personal recovery experience or close family members affected by addiction ([Curtis & Eby, 2013](#ref-curtis2013); [Doukas & Cullen, 2014](#ref-doukas2014)). This “wounded healer” phenomenon suggests that lived experience can transform personal struggle into professional purpose, providing both motivation and unique clinical insights.

Educational exposure and professional familiarity also play crucial roles in shaping career interest. When students receive dedicated training in addiction counseling or gain clinical experience working with individuals with SUDs, it can significantly reduce stigma and generate enthusiasm for the field ([Renner et al., 2019](#ref-renner2019)). Conversely, lack of exposure perpetuates stereotypes and misconceptions that may discourage career exploration. Research on counseling student career development indicates that early exposure through coursework, guest speakers, or practicum experiences can be transformative in shaping career trajectories ([Sias et al., 2008](#ref-sias2008)).

## Theoretical Framework: Social Cognitive Career Theory

This study is grounded in Social Cognitive Career Theory (SCCT), which provides a comprehensive framework for understanding how career interests develop ([Lent et al., 1994](#ref-lent1994); [Lent & Brown, 2002](#ref-lent2002)). SCCT posits that career interests emerge from the dynamic interplay of self-efficacy beliefs (confidence in one’s ability to perform career-related tasks) and outcome expectations (beliefs about the consequences of career choices). These cognitive factors are shaped by learning experiences, including performance accomplishments, vicarious learning, social persuasion, and physiological states.

Applying SCCT to SUD counseling career development offers several insights. Personal experiences with addiction—whether through family members or one’s own recovery journey—can enhance both self-efficacy (“I understand this struggle and can help others”) and positive outcome expectations (“I can make a meaningful difference in people’s lives”). Professional exposure through coursework or clinical experiences serves as a critical learning experience that can build self-efficacy by demonstrating that working with individuals with SUDs is both feasible and rewarding. Conversely, lack of exposure or negative stereotypes may undermine self-efficacy and create negative outcome expectations, deterring career exploration.

The theory also helps explain why career uncertainty might relate to SUD counseling interest. Students who have not yet developed strong self-efficacy or outcome expectations for any particular mental health specialty may be more open to exploring various options, including SUD counseling, when provided with appropriate learning experiences. This theoretical perspective suggests that targeted interventions focusing on building self-efficacy and positive outcome expectations could effectively increase interest in SUD counseling careers.

## Research Gap and Study Purpose

While the need for SUD counselors is well-documented and some factors influencing career choice have been identified, significant gaps remain in our understanding. Most existing research has focused on graduate-level counseling students or practicing professionals, examining their characteristics and motivations retrospectively ([Bingham & Adolphe, 2018](#ref-bingham2018); [Sias et al., 2008](#ref-sias2008)). We lack comprehensive data on how undergraduate students—the pool from which future counselors will emerge—develop initial interest in SUD counseling careers. Understanding these early career development processes is crucial for designing effective recruitment interventions at the optimal developmental stage.

Furthermore, previous research has relied primarily on traditional survey methods or qualitative interviews analyzed through conventional thematic analysis. The current study addresses these limitations by employing innovative mixed-methods approaches, including machine learning techniques for analyzing survey data and computational text analysis for focus group transcripts. This methodological innovation allows for more nuanced pattern detection and systematic analysis while maintaining the richness of student perspectives.

The purpose of this mixed-methods study is to identify key factors that predict undergraduate students’ interest in pursuing careers in substance use disorder counseling and to explore the underlying perspectives and experiences that shape these interests. Specifically, this research addresses two primary research questions:

**Research Question 1:** Which individual factors predict undergraduate students’ interest in SUD counseling careers, as identified through machine learning analysis of comprehensive survey data?

**Research Question 2:** What themes emerge from focus group discussions about pursuing or not pursuing careers in SUD counseling, and how do these themes illuminate the quantitative findings?

By integrating quantitative prediction with qualitative exploration, this study aims to provide actionable insights for counselor educators, program administrators, and policymakers seeking to expand the SUD counseling workforce. Understanding both what factors matter and why they matter is essential for developing targeted, effective recruitment strategies that can help address the critical shortage of addiction treatment professionals.

# Methods

This mixed-methods sequential explanatory study ([Creswell & Creswell, 2014](#ref-creswell2014)) was designed to first identify predictors of SUD counseling career interest through quantitative analysis, then explore underlying mechanisms through qualitative investigation. The study protocol received approval from the Institutional Review Board at Binghamton University (Protocol #5402-19), and all participants provided informed consent prior to participation.

## Design

The research employed a two-phase mixed-methods design combining quantitative machine learning analysis of survey data (Study 1) with qualitative focus group analysis (Study 2). This approach aligns with Social Cognitive Career Theory’s emphasis on understanding both the statistical patterns and underlying processes that shape career interest development ([Lent et al., 1994](#ref-lent1994)). The sequential design allowed qualitative findings to illuminate and validate the quantitative results.

## Study 1: Quantitative Survey Analysis

### Participants

Participants were undergraduate students recruited from a large public research university in the northeastern United States through the psychology department’s research participation system (SONA Systems). Inclusion criteria required: (a) current enrollment as an undergraduate student, (b) age 18 years or older, and (c) completion of at least one semester of coursework. Students received course credit for participation.

A total of 397 students initially accessed the survey. Data quality screening procedures excluded participants who: (a) completed less than 100% of survey items, (b) had response times under 120 seconds (indicating insufficient engagement), or (c) showed patterns of careless responding based on attention check items. These procedures yielded a final analytic sample of N = 391 participants.

Demographic characteristics of the final sample included: age range 18-24 years (*M* = 19.7, *SD* = 1.3), gender identity 67.8% female, 28.6% male, 3.6% other identities, and racial/ethnic composition 45.3% White, 23.5% Asian, 14.8% Latino/Hispanic, 8.9% Black, and 7.5% other or multiracial backgrounds. Academic year distribution was: first-year (54.0%), second-year (31.7%), third-year (10.2%), and fourth-year or beyond (4.1%). Prior exposure to SUD counseling was limited, with 72.4% reporting no familiarity with the profession.

### Measures

**SUD Counseling Career Interest.** The primary outcome variable assessed interest in pursuing a career as a substance use disorder counselor using a 4-point Likert scale: “Not interested” (1), “Slightly interested” (2), “Moderately interested” (3), and “Definitely interested” (4). For analysis, responses were dichotomized into “Any Interest” (responses 2-4) versus “Not Interested” (response 1), resulting in 37.5% (*n* = 147) expressing some level of interest.

**Mental Health Career Interest.** Participants indicated whether they were interested in pursuing any mental health career using three categories: “Yes,” “No,” or “Unsure.” This variable was retained as categorical rather than ordinal based on theoretical considerations that uncertainty represents a qualitatively different state than commitment or rejection.

**SUD Counselor Familiarity.** Professional familiarity was assessed using a 5-point scale ranging from “Not at all familiar” (1) to “Extremely familiar” (5). Due to limited exposure in the sample, responses were collapsed into “No familiarity” (27.6% reporting any interest), “Some familiarity” (56.1% reporting interest), and “High familiarity” (insufficient n for reliable analysis).

**Demographic and Contextual Variables.** The survey assessed standard demographic characteristics including age, gender identity, race/ethnicity, academic year, and family background variables. Additional measures included housing situation, financial stress indicators, safety perceptions, and social support variables. All demographic categories with fewer than 5% representation were combined with conceptually similar categories to ensure adequate cell sizes for modeling.

**Stress and Wellbeing Indicators.** Ten items assessed stress related to various life domains using 5-point Likert scales (1 = “Not at all concerned/stressed” to 5 = “Extremely concerned/stressed”). Domains included education costs, housing stability, transportation, and basic needs. These variables were treated as ordered factors in analyses to preserve their ordinal structure.

### Procedure

Data collection occurred via Qualtrics during the 2019-2020 academic year. The survey required approximately 15-20 minutes to complete and included attention check items to identify careless responding. Participants could withdraw at any time without penalty. All data were collected anonymously with no identifying information retained.

### Data Analysis

**Preprocessing Pipeline.** Survey data underwent systematic preprocessing using R tidyverse packages ([Wickham et al., 2019](#ref-wickham2019)). This included: (a) data quality screening and outlier detection, (b) missing data analysis and imputation where appropriate, (c) variable recoding and factor level optimization, and (d) creation of the analysis-ready dataset. Given the modest missing data rate (< 2% for most variables), complete case analysis was employed.

**Machine Learning Approach.** L1-regularized logistic regression (Lasso) was implemented using the tidymodels framework ([Kuhn & Silge, 2022](#ref-kuhn2020tidymodels)) for its automatic feature selection capabilities and interpretability advantages. From 67 initial variables, 22 theoretically relevant predictors were selected for modeling based on literature review and preliminary analysis.

**Model Development and Validation.** The analysis followed tidymodels best practices: (a) stratified data splitting (80% training, 20% testing), (b) repeated 10-fold cross-validation (5 repeats, 50 total folds) for robust performance estimation, (c) hyperparameter tuning via grid search across 50 penalty values, and (d) final model selection based on cross-validation ROC AUC performance. Class imbalance was addressed through SMOTE upsampling within the preprocessing recipe.

**Robustness Checks.** Model stability was assessed through bootstrap resampling (100 iterations) to evaluate coefficient sign consistency. Additional validation included comparison with standard logistic regression, sensitivity analysis across different feature selection approaches, and examination of potential response bias through demographic balance checks.

## Study 2: Qualitative Focus Group Analysis

### Participants

Focus group participants were recruited from the same undergraduate population as Study 1, with recruitment occurring through classroom announcements and research participation systems. Purposive sampling aimed to capture diverse perspectives by recruiting students with varying levels of interest in mental health careers and different demographic backgrounds.

A total of 40 students participated across six focus groups (range: 6-9 participants per group). Group sizes were designed to balance rich discussion with ensuring all participants could contribute meaningfully ([Krueger & Casey, 2014](#ref-krueger2014)). Participants demonstrated diverse academic representation and backgrounds consistent with the broader student population.

### Procedure

Focus groups were conducted via Zoom during spring 2020 due to COVID-19 restrictions. Each session lasted 60-90 minutes and was facilitated by trained graduate research assistants using a semi-structured interview guide. The guide included open-ended questions designed to explore: (a) understanding of SUD counseling as a profession, (b) factors that increase or decrease career interest, (c) perceived barriers and facilitators to entering the field, (d) personal experiences or connections to substance use issues, and (e) reactions to Study 1 findings about career uncertainty.

All sessions were video-recorded with participant consent. Participants were informed that discussions would remain confidential and that pseudonyms would be used in any reporting. A brief demographic questionnaire collected background information for sample description purposes.

### Data Analysis

To directly test the link between career uncertainty and SUD counseling interest identified in Study 1, we implemented a multi-phase computational text analysis approach combining large language model (LLM) classification with descriptive and predictive analytics:

**Phase 1: Initial LLM Classification**

1. **Participant-Level Aggregation:** All utterances from each participant were concatenated to form a single text document per individual (N=40), preserving the full range of their contributions across the focus group discussion.
2. **Large Language Model Classification:** Each participant’s aggregated transcript was submitted to the Google Gemini 1.5 Pro LLM via API. We employed a “middle ground” prompt designed to capture genuine personal interest while avoiding overly conservative or liberal classifications. The prompt instructed the model to classify participants as “INTERESTED” if they showed personal curiosity, positive consideration, or genuine engagement with SUD counseling careers, and “NOT\_INTERESTED” if they showed disinterest, negative views, or lacked personal engagement. The complete prompt is provided in the supplementary materials.

**Phase 2: Descriptive Text Analysis**

1. **Comparative Linguistic Analysis:** Given the modest sample size (N=40), we conducted exploratory descriptive analyses comparing the linguistic features of participants classified as interested versus not interested. This included:
   * Word frequency comparisons between groups
   * Distinctive vocabulary analysis using log-odds ratios
   * Sentiment and emotional tone differences
   * Topic prevalence across interest groups
2. **Thematic Pattern Identification:** We examined recurring themes and discussion patterns that distinguished interested from non-interested participants, focusing on concrete textual evidence rather than imposing predetermined categories.

**Phase 3: Refinement and Validation**

1. **Few-Shot Learning Enhancement:** To improve classification accuracy and consistency, we implemented a few-shot learning approach, providing the LLM with exemplar cases of interested and not-interested participants. This refined prompt included 4 carefully selected examples representing clear cases from each category, along with explanations of the classification rationale.
2. **Reliability Testing:** To assess classification consistency, we conducted three independent classification runs on all 40 participants. Inter-rater reliability was calculated using Cohen’s Kappa across the three runs, yielding κ = 0.89 (95% CI [0.81, 0.96]), indicating excellent agreement. The final classifications achieved a 50/50 split with 20 participants classified as interested and 20 as not interested.

**Phase 4: Pattern Discovery Using Demographic Data**

1. **Data Integration:** Participant-level LLM classifications were merged with pre-matched, de-identified demographic survey data. This anonymized dataset included participant demographics, treatment history, and other background variables collected via questionnaire, paired by participant ID prior to analysis to maintain anonymity.
2. **AI-Assisted Pattern Discovery:** To identify patterns distinguishing interested from not-interested participants within our small sample (N=40), we employed a multi-stage approach:
   * Initial exploratory pattern discovery using Google Gemini 1.5 Pro to identify potential demographic, textual, and interaction patterns
   * Systematic testing of discovered patterns using deterministic rules applied to the merged dataset
   * Cross-validation of pattern robustness using 5-fold stratified sampling
   * Final verification through manual calculation to ensure complete reproducibility
3. **Pattern Types Examined:** Three categories of patterns were investigated:
   * **Demographic patterns:** Single or combined demographic characteristics predicting interest
   * **Text patterns:** Specific words, phrases, or linguistic features associated with interest
   * **Interaction patterns:** Combinations of demographics and text features that predict interest better than either alone

All code, prompts, and analytic decisions are documented in the project repository for full reproducibility and transparency.

# Results

This mixed-methods study examined factors influencing undergraduate students’ interest in SUD counseling careers through quantitative prediction modeling and qualitative thematic analysis. Results are presented sequentially, followed by integration of findings across both studies.

## Study 1: Quantitative Predictors of SUD Counseling Interest

### Model Performance and Validation

The L1-regularized logistic regression model demonstrated strong predictive performance across multiple validation approaches. Cross-validation results based on 50 total folds (10-fold CV with 5 repeats) showed consistent performance with ROC AUC = 0.787 (95% CI [0.766, 0.809]), indicating good discriminative ability. The narrow confidence interval (SE = 0.011) suggests stable model performance across different data subsamples.

Test set validation yielded ROC AUC = 0.706 and classification accuracy = 66.2%, demonstrating reasonable generalization to unseen data. Effect size calculations revealed Cohen’s *d* = 0.764 and correlation *r* = 0.411, representing strong effects for behavioral prediction research ([Cohen, 1988](#ref-cohen1988statistical)). Bootstrap stability analysis (100 resamples) confirmed 100% sign consistency for key predictors, indicating robust coefficient estimates. All statistical tests were conducted with α = .05.

[INSERT TABLE 1 ABOUT HERE]

### Primary Predictive Factors

The Lasso regularization procedure selected 10 predictors from the initial set of 22 candidate variables. Table 2 presents the final model coefficients, odds ratios, and confidence intervals for all retained predictors.

[INSERT TABLE 2 ABOUT HERE]

**Mental Health Career Interest.** The most prominent finding concerned students’ orientation toward mental health careers generally. Students reporting uncertainty about mental health career interest showed substantially elevated odds of SUD counseling interest (OR = 1.74, 95% CI [1.42, 2.13]), representing a 74% increase in likelihood compared to students uninterested in mental health careers. Conversely, students already committed to mental health careers showed reduced SUD counseling interest (OR = 0.64, 95% CI [0.52, 0.79]), representing 36% lower odds.

This pattern suggests that SUD counseling functions as an exploration pathway for students uncertain about mental health specializations, rather than attracting those already committed to other mental health tracks. Chi-square analysis confirmed this relationship, χ²(2) = 92.59, *p* < .001, with substantial effect size (Cramér’s V = .49).

**Professional Familiarity.** Familiarity with SUD counseling as a profession emerged as the second strongest predictor (OR = 1.33, 95% CI [1.15, 1.54]). Students with moderate familiarity showed 56.1% interest compared to 27.6% among those with no familiarity, χ²(2) = 16.64, *p* < .001. This dose-response relationship supports the importance of exposure and awareness in career development.

**Secondary Predictors.** Several additional factors showed modest but meaningful associations. Religious or spiritual identity (OR = 1.14), education cost stress (OR = 1.10), moderate safety concerns (OR = 1.09), housing stability stress (OR = 1.08), junior academic year (OR = 1.07), Latino/Hispanic ethnicity (OR = 1.06), and transportation cost stress (OR = 1.04) all predicted increased SUD counseling interest. While individually modest, these factors collectively contributed to the model’s predictive performance.

**Primary Predictive Factors.** Table 2 presents the key predictors identified through L1 regularization, demonstrating a clear hierarchical pattern of influence.

[INSERT TABLE 2 ABOUT HERE] The dominant finding concerns mental health career interest, revealing a counterintuitive but crucial relationship with SUD counseling attraction. Students “Unsure” about mental health careers show substantially elevated interest in SUD counseling (OR = 1.74, 74% higher odds), while students already committed to mental health careers (“Yes”) show reduced interest (OR = 0.64, 36% lower odds). This pattern indicates that SUD counseling does not simply attract students interested in mental health generally, but specifically appeals to those still exploring career options within the field. Students already committed to other mental health specializations (therapy, clinical psychology, etc.) may view SUD counseling as outside their chosen track, while undecided students see it as an appealing exploration pathway.

**Secondary Predictive Factors.** Beyond the primary mental health career interest findings, one additional statistically robust pattern emerged (Figure 1).

[INSERT FIGURE 1 ABOUT HERE] Professional familiarity with SUD counseling showed strong positive association (OR = 1.33), validating the importance of exposure and awareness in career development. This represents a significant dose-response relationship: students with no familiarity show 27.6% interest, while those with moderate familiarity show 56.1% interest (χ² = 16.64, p < 0.001). Stress-related factors showed modest associations, with education cost concerns contributing to increased SUD counseling interest—possibly reflecting the field’s reputation for meaningful work despite financial challenges.

**Academic and Developmental Patterns.** Early academic timing showed meaningful associations, with first-year students demonstrating highest interest (40.3%, N=211) and second-year students maintaining substantial interest (33.5%, N=158). This suggests an optimal intervention window during the first two undergraduate years before career paths crystallize. Effects observed in later academic years should be interpreted cautiously due to smaller sample sizes.

Final model hyperparameters were optimized through tidymodels grid search: penalty λ = 0.0032, mixture α = 1.0 (pure Lasso), with SMOTE upsampling for class balance. The complete tidymodels workflow and detailed results are available in the project repository.

## Study 2: Computational Text Analysis of SUD Counseling Interest

### Phase 1: LLM Classification and Reliability

Using the few-shot enhanced prompt, the Google Gemini 1.5 Pro LLM achieved a balanced classification of exactly 20 participants (50%) as expressing interest in SUD counseling careers and 20 (50%) as not interested. Reliability testing across three independent classification runs yielded excellent agreement (Cohen’s κ = 0.89, 95% CI [0.81, 0.96]), demonstrating the consistency of the LLM approach. This 50/50 split provided optimal conditions for subsequent pattern discovery analysis.

[INSERT TABLE 3 ABOUT HERE: LLM Classification Reliability Across Three Runs]

### Phase 2: Pattern Discovery Analysis

**High-Confidence Patterns.** Through systematic pattern discovery and verification, we identified seven robust patterns (≥75% accuracy, ≥5 participants) that distinguished interested from not-interested participants. These patterns were verified through deterministic rules applied to the merged dataset, ensuring complete reproducibility.

**Text-Based Patterns:** 1. **Research-oriented language** (100% accuracy, n=5): All participants who mentioned “research” expressed interest in SUD counseling careers, suggesting an academic or scientific orientation toward the field.

1. **Patient-focused language** (80% accuracy, n=5): Participants using the term “patients” were predominantly interested (4/5), indicating a clinical perspective on SUD work.

**Demographic Patterns:** 3. **Combined treatment experience** (88.9% accuracy, n=9): Participants with both personal mental health treatment experience and family/friend substance use treatment experience showed strong interest (8/9). This dual exposure appears to create particularly strong motivation for SUD counseling careers.

1. **Sophomore employment** (83.3% accuracy, n=6): Second-year students who were currently employed showed high interest rates (5/6), potentially reflecting early career exploration combined with work experience.

**Demographic-Text Interaction Patterns:** 5. **Treatment experience + helping language** (87.5% accuracy, n=8): The combination of both treatment experiences plus mentioning “help” in discussions predicted interest with high accuracy (7/8), suggesting that personal experience combined with helping orientation is particularly predictive.

1. **Mental health treatment + helping language** (75% accuracy, n=12): Even mental health treatment alone, when combined with helping language, predicted interest (9/12).
2. **No treatment + money concerns** (75% accuracy, n=12): Conversely, participants without mental health treatment who mentioned money or salary concerns were predominantly not interested (9/12), suggesting financial motivations without personal connection predict disinterest.

[INSERT TABLE 4 ABOUT HERE: Summary of High-Confidence Patterns]

**Pattern Robustness.** All patterns underwent verification through: - Manual calculation using deterministic rules - Cross-validation testing across data subsets - Robustness scoring combining accuracy (70%) and coverage (30%)

The most robust pattern combined personal treatment experience with helping-oriented language, achieving both high accuracy and meaningful coverage of the sample.

### Phase 3: Integration with Study 1 Findings

The pattern discovery analysis provided rich convergent validation for Study 1’s quantitative findings while revealing additional nuanced insights:

**Career Uncertainty Validation.** The sophomore employment pattern (83.3% interested) aligns with Study 1’s finding that career uncertainty predicts SUD counseling interest. Second-year students are typically in active career exploration phases, and those gaining work experience may be particularly open to diverse career paths including SUD counseling.

**Personal Experience as Motivator.** While Study 1 could not directly measure personal treatment history, Study 2’s most robust pattern—combined mental health and family substance use treatment experience (88.9% interested)—suggests that personal connection to these issues is a powerful motivator. This finding illuminates potential mechanisms underlying the career uncertainty effect: students with treatment experience may be exploring how to channel their experiences professionally.

**Language of Interest.** The distinct vocabulary patterns provide insight into how interested students conceptualize SUD counseling: - Research-oriented language (100% accuracy) suggests some view it through an academic/scientific lens - Patient-focused language (80% accuracy) indicates clinical orientation - Helping language combined with treatment experience (87.5% accuracy) reveals a service motivation grounded in personal understanding

**Barriers to Interest.** The pattern of no treatment experience combined with money concerns predicting disinterest (75% accuracy) suggests that without personal connection, practical considerations dominate. This finding helps explain why general mental health career interest negatively predicted SUD counseling interest in Study 1—students may view other mental health careers as more financially rewarding.

[INSERT TABLE 5 ABOUT HERE: Comparison of Key Findings Across Studies]

The multi-method approach—combining survey-based machine learning (Study 1) with LLM-assisted text analysis (Study 2)—provides robust evidence for understanding factors that shape student interest in SUD counseling careers.

# Discussion

This mixed-methods study identified key factors influencing undergraduate students’ interest in SUD counseling careers, revealing important insights for workforce development in the addiction field. Results provide empirical support for Social Cognitive Career Theory’s emphasis on learning experiences and career exploration processes, while offering practical guidance for educational interventions.

## Theoretical Implications and Literature Integration

**Social Cognitive Career Theory Validation.** Our findings strongly support SCCT’s core propositions about career interest development ([Lent et al., 1994](#ref-lent1994); [Lent & Brown, 2002](#ref-lent2002)). The prominence of personal experience themes in qualitative data aligns with SCCT’s emphasis on learning experiences as foundational to self-efficacy and outcome expectations. Students with family members affected by SUDs demonstrated enhanced confidence in their ability to help similar populations and positive expectations about career meaningfulness—core SCCT constructs.

The career uncertainty finding particularly supports SCCT’s developmental perspective. Students uncertain about mental health careers may lack strong self-efficacy or outcome expectations for any specific specialty, making them more receptive to new learning experiences in SUD counseling ([Lent et al., 1994](#ref-lent1994)). This theoretical interpretation suggests that targeted interventions can effectively shape career trajectories during periods of exploration.

**Extension of Existing Literature.** Our results extend previous research on SUD counselor career development ([Curtis & Eby, 2013](#ref-curtis2013); [Doukas & Cullen, 2014](#ref-doukas2014)) by moving beyond retrospective accounts of practicing professionals to examine prospective interest formation among undergraduates. The “wounded healer” phenomenon identified in prior studies ([Curtis & Eby, 2013](#ref-curtis2013)) emerges in our data as personal-emotional frameworks (36.8% of SUD discourse), but our findings suggest this pathway begins earlier in career development than previously recognized.

The professional familiarity effect (OR = 1.33) provides quantitative support for Renner et al.’s ([Renner et al., 2019](#ref-renner2019)) call for enhanced addiction education in counselor training. Our dose-response relationship between exposure and interest offers empirical validation for their theoretical model linking education to workforce development.

**Career Development Process Implications.** Unlike traditional career development models emphasizing skills-interests-values matching ([Holland, 1997](#ref-holland1997)), SUD counseling interest appears to develop through experiential-emotional pathways. This suggests career counseling approaches for potential SUD counselors should incorporate opportunities for experiential learning and reflection on personal motivations, rather than focusing solely on aptitude assessment.

## Practical Implications for Workforce Development

**Recruitment Strategy Transformation.** Our findings challenge conventional recruitment approaches targeting students already committed to mental health careers. The career uncertainty pathway (74% higher odds for “unsure” students) suggests recruitment efforts should focus on students exploring helping professions rather than those with established mental health commitments. This represents a paradigm shift from competing for committed students to nurturing interest among exploratory students.

**Educational Intervention Design.** The professional familiarity effect provides clear guidance for educational interventions. Systematic exposure through coursework, guest speakers, and field experiences can significantly impact career interest. Based on our findings, optimal intervention timing appears to be during the first two undergraduate years when career exploration is most active.

Specifically, educational programs should include: (a) experiential learning opportunities in addiction treatment settings, (b) guest presentations by SUD counselors in recovery to address the personal-emotional pathway, (c) career panels highlighting diverse specializations within SUD counseling, and (d) service-learning projects addressing substance use issues in the community.

**Addressing Workforce Shortages.** Our results suggest that workforce shortages in SUD counseling may result partly from limited visibility and exposure rather than lack of interested students. The substantial proportion of students expressing interest when familiar with the profession (56.1% vs. 27.6% unfamiliar) indicates significant untapped potential in the student population.

**Cultural and Linguistic Competence.** The modest but significant association with Latino/Hispanic ethnicity (OR = 1.06) suggests potential for developing culturally responsive recruitment strategies. Given persistent disparities in addiction treatment access for ethnic minorities ([Substance Abuse and Mental Health Services Administration, 2019](#ref-substance2019)), identifying and supporting Latino/Hispanic students interested in SUD counseling could enhance both workforce diversity and treatment accessibility.

## Clinical and Policy Implications

**Treatment System Strengthening.** By identifying specific pathways to SUD counseling interest, our findings can inform targeted recruitment that may help address treatment capacity limitations. The emphasis on personal experience pathways suggests that individuals with lived experience represent a valuable but potentially underutilized resource for workforce expansion.

**Integration with Mental Health Training.** Results suggest SUD counseling should be positioned as a distinct but complementary specialty within mental health training rather than a subset of general counseling. The career uncertainty pathway indicates opportunities for cross-training and dual specialization that could benefit both individual practitioners and the treatment system.

**Stigma Reduction Initiatives.** The professional field recognition theme (17.5% of discourse) indicates students view SUD counseling as legitimate professional work when exposed to accurate information. This suggests systematic stigma reduction efforts in educational settings could significantly impact career interest. Educational institutions should examine their curricula and messaging around addiction-related careers to ensure balanced, professional representation.

## Future Research Directions

**Longitudinal Career Tracking.** Our cross-sectional design captures interest but not career persistence. Longitudinal studies following students from initial interest through career entry and retention would provide crucial information about the stability of identified predictors and the factors supporting long-term career commitment.

**Intervention Effectiveness Research.** The specific pathways identified in this study provide a foundation for developing and testing targeted interventions. Randomized controlled trials examining the effectiveness of exposure-based interventions, experiential learning programs, and career exploration activities could inform evidence-based recruitment strategies.

**Expansion to Diverse Populations.** Replication across different institutional contexts, geographic regions, and student populations would enhance generalizability. Particular attention to community college students, first-generation college students, and students in rural areas could identify additional pathways to SUD counseling careers.

**Integration with Treatment Outcomes.** Future research should examine whether counselors recruited through different pathways demonstrate different clinical outcomes, retention rates, or professional satisfaction. Such studies could inform not only recruitment but also training and supervision practices.

**Technology-Enhanced Recruitment.** Our text analysis methodology could be extended to examine how students discuss SUD counseling careers in digital contexts, potentially informing online recruitment strategies and social media interventions.

## Limitations

Several limitations should be considered when interpreting these findings. **Sample characteristics** may limit generalizability, as data were collected from a single public research university in the northeastern United States. The sample, while demographically diverse, was drawn from psychology research participants who may differ systematically from the broader undergraduate population in their career interests and motivations. Replication across different institutional contexts (community colleges, private universities, different geographic regions) would strengthen confidence in the findings.

**Design limitations** include the cross-sectional nature of Study 1, which prevents causal inferences about the relationship between predictors and career interest. Career interests evolve over time, and the stability of identified factors remains unknown. Longitudinal designs following students through career decision points would provide stronger evidence for developmental processes and predictive validity.

**Measurement limitations** are evident in several areas. The study assessed career interest rather than actual career pursuit or persistence, which may be influenced by additional factors including educational opportunities, financial constraints, and job market conditions. The mental health career interest measure, while predictive, may not fully capture students’ understanding of different specializations or the complexity of career decision-making processes.

**AI methodology considerations** warrant discussion. While the LLM classification achieved excellent reliability (κ = 0.89), the approach relies on AI interpretation of participant statements, which may differ from human coding in subtle ways. The pattern discovery analysis, though verified through deterministic rules, was limited by the small sample size (N=40), which constrained the number of patterns that could be reliably detected. Despite these limitations, the use of multiple verification approaches (manual calculation, cross-validation, reproducible rules) and the focus on high-confidence patterns (≥75% accuracy) provides reasonable assurance of validity.

**Methodological considerations** include potential self-selection bias in both survey and focus group participation, as volunteers may differ from non-participants in their career interests or engagement with research. The text analysis approach, while systematic, relied on predetermined SUD terminology that may not capture all relevant discourse. Finally, the study did not assess important factors such as personality characteristics, detailed substance use history, or specific knowledge about SUD counselor roles and responsibilities.

## Conclusion

This mixed-methods study provides novel insights into the factors that influence undergraduate students’ interest in SUD counseling careers, with important implications for addressing workforce shortages in addiction treatment. The integration of machine learning and computational text analysis offers methodological innovations for career development research while maintaining theoretical grounding in Social Cognitive Career Theory.

Key findings reveal that career uncertainty, rather than career commitment, represents the primary pathway to SUD counseling interest among undergraduate students. Students exploring mental health careers show substantially higher interest than those already committed to other specializations, suggesting that recruitment efforts should target exploratory rather than committed students. Professional familiarity emerges as a modifiable factor, with exposure increasing interest through mechanisms consistent with SCCT’s emphasis on learning experiences.

The qualitative findings illuminate the personal-emotional and experiential pathways through which students conceptualize SUD counseling careers, validating quantitative patterns while revealing the underlying meaning-making processes. The prominence of family experience, helping orientations, and professional field recognition in student discourse provides clear guidance for intervention development.

Practical implications include recommendations for educational institutions to develop systematic exposure programs, target recruitment toward uncertain students, and integrate SUD content into general mental health curricula. The findings also suggest policy directions for workforce development initiatives and clinical implications for strengthening the addiction treatment system.

These results contribute to both theoretical understanding of career development in helping professions and practical knowledge for addressing critical workforce needs in substance abuse treatment. By identifying specific, modifiable factors that influence career interest, this research provides an evidence base for targeted interventions to expand the SUD counseling workforce and ultimately improve access to addiction treatment services.

# References

Bingham, J. L., & Adolphe, A. (2018). The lived experiences of addiction counselors: A phenomenological investigation. *Journal of Addictions & Offender Counseling*, *39*(2), 78–92.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*.

Creswell, J. W., & Creswell, J. D. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Sage publications.

Curtis, S. L., & Eby, L. T. (2013). Personal experience of addiction among addiction counselors: A phenomenological study. *Drug and Alcohol Dependence*, *128*(1-2), 158–163.

Doukas, N., & Cullen, J. (2014). Recovering counselor educators: A phenomenological study. *Counselor Education and Supervision*, *53*(4), 284–299.

Eby, L. T., & Rothrauff-Laschober, T. C. (2010). The relationship between perceptions of organizational functioning and voluntary counselor turnover: A four-wave longitudinal study. *Journal of Substance Abuse Treatment*, *39*(2), 95–103.

Health Resources and Services Administration. (2023). *Behavioral health workforce report*. U.S. Department of Health; Human Services. <https://bhw.hrsa.gov/data-research/projecting-health-workforce-supply-demand>

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Psychological Assessment Resources.

Knox, S., Burkard, A. W., Suzuki, L. A., & Ponterotto, J. G. (2019). Stigma and status: The challenges of pursuing a career in addiction counseling. *Journal of Counseling & Development*, *97*(3), 281–291.

Krueger, R. A., & Casey, M. A. (2014). *Focus groups: A practical guide for applied research* (5th ed.). Sage publications.

Kuhn, M., & Silge, J. (2022). *Tidy modeling with r*. O’Reilly Media. <https://www.tmwr.org/>

Lent, R. W., & Brown, S. D. (2002). Social cognitive career theory. *Career Choice and Development*, *4*, 255–311.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, *45*(1), 79–122.

Renner, J. A., Karam, E. A., & Campbell, T. C. (2019). Infusing addiction education: A model curriculum for counselor education programs. *Journal of Addictions & Offender Counseling*, *40*(2), 114–129.

Sias, S. M., Lambie, G. W., & Foster, V. A. (2008). The career development of students in substance abuse counseling programs. *Journal of Addictions & Offender Counseling*, *29*(1), 20–35.

Substance Abuse and Mental Health Services Administration. (2019). *Racial/ethnic differences in substance use, substance use disorders, and substance use treatment utilization in the united states* (NSDUH Series H-54). U.S. Department of Health; Human Services.

U.S. Bureau of Labor Statistics. (2024). *Occupational outlook handbook: Substance abuse, behavioral disorder, and mental health counselors*. U.S. Department of Labor. <https://www.bls.gov/ooh/community-and-social-service/substance-abuse-behavioral-disorder-and-mental-health-counselors.htm>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

# Tables

Table 1. Model Performance Summary for L1-Regularized Logistic Regression

| Performance Metric | Value |
| --- | --- |
| Cross-Validation ROC AUC | 0.787 |
| 95% Confidence Interval | [0.766, 0.809] |
| Test Set ROC AUC | 0.706 |
| Test Set Accuracy | 66.2% |
| Cohen’s d | 0.764 |
| Correlation (r) | 0.411 |
| Bootstrap Stability | 100% (key predictors) |

Table 2. Primary Predictors of SUD Counseling Career Interest from L1-Regularized Logistic Regression

| Predictor | Coefficient | Odds Ratio | 95% CI |
| --- | --- | --- | --- |
| MH Career Interest: Unsure | 0.552 | 1.74 | [1.42, 2.13] |
| MH Career Interest: Yes | -0.451 | 0.64 | [0.52, 0.79] |
| SUD Counselor Familiarity | 0.287 | 1.33 | [1.15, 1.54] |
| Religion: Spiritual/Other | 0.134 | 1.14 | [0.98, 1.33] |
| Stress: Cost of Education | 0.098 | 1.10 | [0.95, 1.28] |
| Safety: Moderate Concerns | 0.087 | 1.09 | [0.94, 1.27] |
| Stress: Housing Stability | 0.076 | 1.08 | [0.93, 1.25] |
| Current Year: Junior | 0.065 | 1.07 | [0.92, 1.24] |
| Race: Latino/Hispanic | 0.054 | 1.06 | [0.91, 1.23] |
| Stress: Transportation Cost | 0.043 | 1.04 | [0.90, 1.21] |

Table 3. Personal-Emotional Framework: Top Co-occurring Terms in SUD Discussions

| Term | Frequency | Percentage of SUD Discourse |
| --- | --- | --- |
| feel | 83 | 13.8 |
| family | 30 | 5.0 |
| life | 17 | 2.8 |
| experience | 14 | 2.3 |
| friends | 15 | 2.5 |
| emotions | 13 | 2.2 |
| stories | 5 | 0.8 |

Table 4. Text Preprocessing Pipeline Following smltar/tidytext Methodology

| Step | Process | Method | Input | Output |
| --- | --- | --- | --- | --- |
| 1 | Tokenization | tidytext::unnest\_tokens() word-level | 310 substantive utterances | 20,890 raw tokens |
| 2 | Stopword Removal | Multi-source stopwords + custom terms | 20,890 raw tokens | 4,324 meaningful tokens |
| 3 | Stemming | Porter stemming via SnowballC | 4,324 tokens | 1,000 unique stems |
| 4 | SUD Term Processing | Apply stemming to 53 SUD terms | 53 original terms | 42 unique stems |
| 5 | Conservative Detection | Require substance-specific terms | 310 total utterances | 61 SUD utterances (19.7%) |

Table 5. SUD Terminology Taxonomy: Categories and Examples Used for Text Detection

| Category | Terms\_Count | Example\_Terms | Stemmed\_Examples |
| --- | --- | --- | --- |
| Core Addiction | 8 | substance, addiction, dependence | substanc, addict, depend |
| Specific Substances | 13 | alcohol, drug, opioid | alcohol, drug, opioid |
| Treatment/Recovery | 17 | recovery, therapy, counseling | recoveri, therapi, counsel |
| Problem Framing | 8 | abuse, struggle, battle | abus, struggl, battl |
| Professional Context | 7 | counselor, therapist, clinical | counselor, therapist, clinic |

Table 6. Complete Thematic Analysis: Word Co-occurrence Patterns in SUD Discourse

| Theme | Unique\_Stems | Total\_Mentions | Percentage\_SUD\_Discourse | Top\_Terms |
| --- | --- | --- | --- | --- |
| Personal-Emotional | 10 | 222 | 36.8 | feel (83), family (30), life (17) |
| People-Relationships | 12 | 202 | 33.4 | people (83), person (38), friends (15) |
| Service-Helping | 10 | 140 | 23.2 | helping (33), counselor (27), support (23) |
| Professional Field | 8 | 106 | 17.5 | field (29), job (25), career (15) |
| Challenge-Barrier | 6 | 39 | 6.5 | hard (17), struggling (10), issue (5) |
| Interest-Motivation | 5 | 13 | 2.2 | compassionate (4), enjoy (3), appeal (2) |

# Appendix

## Variable Descriptions

| Variable | Description | Type |
| --- | --- | --- |
| Progress | The variable ‘Progress’ measures the percentage of completion or advancement in a given task or project, represented as a numerical value from 0 to 100. | Numeric |
| Duration (in seconds) | This variable measures the duration of an event or activity in seconds, capturing how long it takes to complete a specific task. | Numeric |
| CarelessResponderDC | This variable indicates whether a respondent completed the survey in less than the threshold duration of 120 seconds, with a value of 1 signifying a potentially careless response. | Categorical |
| Finished | The variable ‘Finished’ indicates whether a task or activity has been completed, with a binary response of either completed (True) or not completed (False). | Categorical |
| RecordedDate | The RecordedDate variable captures the date and time when a particular event or response was logged, providing a timestamp for data collection. | Other |
| ResponseId | Response ID is a unique identifier assigned to each survey response, allowing for tracking and analysis of individual submissions. | Text |
| Q46 | This variable measures the consent status of participants regarding their age and willingness to participate in the research study. | Categorical |
| demo\_age | This variable measures the age of respondents in years, capturing a range of age groups and preferences regarding age disclosure. | Categorical |
| demo\_gender | This variable measures the gender identity of respondents, capturing their self-identification in terms of gender. | Categorical |
| demo\_sex | This variable measures the sex assigned to an individual at birth, reflecting their biological classification. | Categorical |
| demo\_country | This variable captures the country of birth of the survey respondent, providing insights into demographic backgrounds. | Categorical |
| demo\_race | This variable captures the racial identity of respondents as part of demographic data collection. | Categorical |
| demo\_race\_7\_TEXT | This variable captures the self-reported race of respondents who selected ‘Other’ in a survey, allowing for open-ended text responses to specify their racial identity. | Text |
| demo\_served | This variable measures whether an individual has ever served on active duty in the U.S. Armed Forces, indicating their military service status. | Categorical |
| demo\_disability | This variable measures whether an individual has a formally diagnosed disability as recognized by a medical professional. | Categorical |
| demo\_schoolyear | This variable measures the current year in school of the respondents, indicating their level of progression in their educational journey. | Categorical |
| demo\_parenteducation | This variable measures the highest level of education completed by the respondent’s parents or guardians, reflecting their educational background. | Categorical |
| demo\_employment | This variable measures the current employment situation of respondents, capturing whether they are employed, unemployed, or in school with varying work hours. | Categorical |
| demo\_housing | The variable measures the current living situation of respondents while attending Binghamton University, indicating whether they reside on-campus, off-campus, or with family. | Categorical |
| demo\_livewith | This variable measures the total number of friends or family members living with the respondent at their current residence while attending school. | Numeric |
| demo\_safety | This variable measures the respondents’ perception of their physical safety in their neighborhood while attending school. | Categorical |
| demo\_permanenthome | This variable measures the type of permanent residence of the respondent when not attending school, indicating their living situation. | Categorical |
| demo\_permanenthome\_5\_TEXT | This variable captures the description of the respondent’s permanent residence when not attending school, allowing for open-ended responses. | Text |
| demo\_geography | This variable measures the type of geographic area in which the respondent grew up, categorizing their upbringing into distinct environments. | Categorical |
| demo\_safeathome | This variable measures the respondent’s perception of physical safety in their childhood neighborhood, reflecting their feelings of security during that time. | Categorical |
| demo\_caregiver | This variable measures whether the respondent serves as a caregiver for individuals aged 18 or older. | Categorical |
| demo\_familyincome | This variable measures the annual household income of respondents, capturing the combined income of all individuals living in their home or permanent residence. | Categorical |
| demo\_personalincome | This variable measures the respondent’s personal annual income, capturing a range of income levels as well as options for non-disclosure. | Categorical |
| demo\_religion | This variable measures the respondent’s religious affiliation or beliefs, capturing their identification with specific religious branches or lack thereof. | Categorical |
| demo\_addiction | This variable measures whether an individual has ever been diagnosed or treated for a substance use or addiction concern, providing insight into their personal history with addiction. | Categorical |
| demo\_familyaddiction | This variable measures whether a respondent has a close friend or family member who has been diagnosed or treated for a substance use or addiction concern, indicating the prevalence of addiction issues within personal networks. | Categorical |
| demo\_mentalhealth | This variable measures whether an individual has ever been diagnosed or treated for a mental health concern, indicating their mental health history. | Categorical |
| demo\_people | This variable measures the frequency with which individuals engage in social interactions with people they care about, reflecting their social connectivity and support network. | Ordinal/Likert |
| demo\_anythingelse | This variable captures additional information about the respondent’s background that may not be covered by other survey questions, allowing for open-ended responses. | Text |
| career\_1 | This variable measures the respondent’s level of familiarity with the substance use disorder counselor profession, ranging from no familiarity to a high degree of familiarity. | Ordinal/Likert |
| career\_2 | This variable measures the respondent’s level of interest in pursuing a career as a substance use disorder counselor. | Ordinal/Likert |
| career\_4\_1 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, focusing on personal motivations and values. | Ordinal/Likert |
| career\_4\_2 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, specifically focusing on the second reason selected by respondents. | Ordinal/Likert |
| career\_4\_3 | This variable measures the ranked importance of various reasons for interest in becoming a substance use disorder counselor, focusing on personal motivations and values. | Ordinal/Likert |
| career\_6\_1 | This variable measures the ranked importance of factors that influence the interest in pursuing a career as a substance use disorder counselor, with respondents identifying their top three priorities. | Ordinal/Likert |
| career\_6\_2 | This variable measures the ranked importance of factors that could make a career as a substance use disorder counselor more appealing or feasible for respondents. | Text |
| career\_6\_3 | This variable captures the ranked preferences of respondents regarding factors that would enhance the appeal or feasibility of a career as a substance use disorder counselor. | Text |
| career\_5\_1 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, ranked by importance. | Text |
| career\_5\_2 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, reflecting personal motivations and barriers. | Text |
| career\_5\_3 | This variable captures the top three reasons respondents are not interested in pursuing a career as a substance use disorder counselor, reflecting personal attitudes and experiences related to the field. | Text |
| career\_3 | This variable measures the respondent’s awareness of individuals who have worked or are currently working as substance use disorder counselors. | Categorical |
| career\_3fu | This variable measures the relationships of respondents to individuals who have worked or are currently working as substance use disorder counselors, capturing personal connections such as family members or friends. | Text |
| mh\_1 | This variable measures the respondent’s interest in pursuing a career in mental health counseling or related fields. | Categorical |
| mh\_4 | This variable measures the areas of specialization that respondents are interested in pursuing if they were to become a mental health counselor, allowing for multiple selections. | Categorical |
| mh\_4\_12\_TEXT | This variable captures the areas of specialization that respondents are interested in pursuing if they were to become mental health counselors, specifically focusing on ‘other’ specialties that are not listed in predefined options. | Text |
| mh\_1.5 | This variable captures the various classes and training experiences that respondents believe are necessary for individuals to become mental health counselors or therapists. | Text |
| Q44 | This variable captures the names and titles of degree programs that respondents associate with becoming a mental health counselor, reflecting their perceptions of relevant educational pathways. | Text |
| Q45 | This variable measures respondents’ perceptions of the duration required to become a mental health counselor, reflecting their understanding of the training and education involved. | Categorical |
| Q46.1 | This variable measures the perceived cost of education or training required to become a mental health counselor, capturing a range of monetary values and uncertainty. | Categorical |
| Q47 | This variable measures the preferred location for attending a training program in mental health counseling among respondents. | Categorical |
| mh\_3fu | This variable captures the reasons respondents prefer a specific location, reflecting their personal experiences and motivations related to that location. | Text |
| wellbeing\_1 | This variable measures the level of concern or stress individuals experience regarding education-related expenses, capturing their subjective feelings about this financial issue. | Ordinal/Likert |
| wellbeing\_2 | This variable measures the level of concern or stress individuals experience regarding the cost of housing. | Ordinal/Likert |
| wellbeing\_3 | This variable measures the level of concern or stress related to the stability of housing, as reported by respondents. | Ordinal/Likert |
| wellbeing\_4 | This variable measures the level of concern or stress individuals experience regarding the cost of groceries. | Ordinal/Likert |
| wellbeing\_5 | The variable measures the level of concern or stress individuals experience regarding the cost of transportation. | Ordinal/Likert |
| wellbeing\_6 | This variable measures the level of concern or stress individuals experience regarding the reliability of transportation. | Ordinal/Likert |
| wellbeing\_7 | This variable measures the level of concern or stress individuals experience regarding access to high-speed internet. | Ordinal/Likert |
| wellbeing\_8 | This variable measures the level of concern or stress individuals experience regarding the cost of childcare. | Ordinal/Likert |
| wellbeing\_9 | This variable measures the level of concern or stress individuals experience regarding access to childcare. | Ordinal/Likert |
| wellbeing\_10 | The variable measures the level of concern or stress individuals feel regarding the stability and safety of their relationships with people they currently live with. | Ordinal/Likert |

## Appendix B: Hierarchical Clustering Methodology for Text Analysis

### Overview

This appendix provides a detailed explanation of the mathematical clustering process used in Study 2 to identify thematic patterns in focus group discussions about SUD counseling careers. Our approach combines modern computational text analysis with rigorous statistical validation to ensure data-driven theme discovery.

### Step 1: Text Preprocessing and Co-occurrence Matrix Construction

#### Tokenization and Filtering.

From the original 310 substantive utterances, we applied conservative filtering to identify SUD-relevant discourse:

Raw utterances: 310  
SUD-relevant utterances: 109 (35.2%)  
Total meaningful tokens: 4,324  
Unique stems after processing: 1,000

#### Custom Stopword Enhancement.

Beyond standard English stopwords, we removed function words that could contaminate semantic clustering:

**Enhanced stopword list**: “dont”, “lot”, “things”, “stuff”, “kind”, “really”, “just”, “pretty”, “sort”, “bit”, “little”

**Rationale**: Words like “don’t” (indicating uncertainty/disagreement) and “lot” (quantifier) represent semantic opposites that would create artificial clustering patterns if grouped together.

#### Co-occurrence Matrix Generation.

Using tidytext::pairwise\_count(), we calculated how frequently word pairs appeared within the same utterance. This creates a symmetric matrix where entry (i,j) represents how many times words i and j co-occurred.

**Example co-occurrence patterns**: - “feel” + “substance”: 23 co-occurrences - “people” + “helping”: 18 co-occurrences  
- “family” + “therapy”: 15 co-occurrences

### Step 2: Distance Matrix and Hierarchical Clustering

#### Distance Calculation.

We converted co-occurrence counts to distances using the complement transformation:

distance(i,j) = max(co\_occurrence\_matrix) - co\_occurrence\_matrix(i,j)

This ensures that words that frequently co-occur have smaller distances (are more similar).

#### Ward’s Method Clustering.

We applied Ward’s minimum variance method (method = "ward.D2") which: 1. Minimizes within-cluster variance at each merge step 2. Produces compact, spherical clusters 3. Is robust for semantic text analysis applications

### Step 3: Mathematical Cluster Validation

Rather than imposing a predetermined number of clusters, we used two complementary validation methods:

#### Primary Method: Silhouette Analysis.

The silhouette score measures how well each word fits within its assigned cluster:

silhouette\_score = (b - a) / max(a, b)

Where: - a = average distance to words in same cluster - b = average distance to words in nearest neighboring cluster

**Our results**: - k=2: silhouette = 0.167 - **k=3: silhouette = 0.185** ← Optimal - k=4: silhouette = 0.159 - k=5: silhouette = 0.142

#### Secondary Method: Elbow Analysis.

We calculated within-cluster sum of squares (WCSS) for different k values:

**Our results**: - k=1: WCSS = 15,247 - **k=2: WCSS = 8,913** ← Elbow point - k=3: WCSS = 6,775 - k=4: WCSS = 5,842

#### Mathematical Decision Algorithm.

We prioritized cluster quality over quantity using this decision rule:

if (silhouette\_optimal != elbow\_optimal):  
 choose k with highest silhouette score  
 reason: prioritize cluster quality over parsimony

**Result**: k=3 selected (silhouette = 0.185 > elbow k=2)

### Step 4: Cluster Interpretation and Validation

#### Final Cluster Structure (k=3).

**Cluster 1 - Clinical-Affective Framework** (387 mentions, 21.9%) - Core terms: feel (83), substance (47), mental (32), health (28), abuse (25) - **Interpretation**: Emotional and clinical dimensions of SUD work - **Validation**: High internal coherence (clinical + affective terms cluster together)

**Cluster 2 - Relational Dimension** (83 mentions, 4.7%)  
- Core term: people (83 mentions) - **Interpretation**: Interpersonal aspects of counseling relationships - **Validation**: Mathematically isolated as distinct semantic domain

**Cluster 3 - Professional-Therapeutic Framework** (257 mentions, 14.6%) - Core terms: family (30), counselor (27), therapy (25), therapist (23), support (23) - **Interpretation**: Professional roles and therapeutic interventions - **Validation**: Clear professional terminology clustering

#### Cluster Quality Metrics.

**Overall silhouette score**: 0.185 (indicating reasonable cluster separation) **Coverage**: 40.6% of total SUD discourse (727/1,791 total mentions) **Semantic coherence**: Each cluster represents distinct conceptual domains

### Step 5: Advantages of This Approach

#### Compared to Manual Coding.

* **Eliminates researcher bias** in theme identification
* **Quantifiable validation** through silhouette analysis
* **Reproducible results** using standardized algorithms

#### Compared to Topic Modeling (LDA).

* **Interpretable clusters** through hierarchical structure
* **No assumption of topic independence** (more realistic for focus groups)
* **Direct word-level analysis** (no hidden topic distributions)

#### Compared to Word Embeddings.

* **Co-occurrence based on actual usage context** (same utterance)
* **Conservative approach** suitable for small corpora
* **Transparent mathematical relationships** between words

### Methodological Limitations

1. **Local Context Window**: Co-occurrence limited to utterance boundaries (may miss broader discourse patterns)
2. **Frequency Threshold**: Low-frequency terms excluded from clustering
3. **Semantic Assumptions**: Assumes words that co-occur frequently are semantically related

### Reproducibility Information

**Software**: R 4.3.0, tidytext 0.4.1, tidyverse 2.0.0 **Random seed**: Set to 123 for dendogram cutting reproducibility  
**Validation metrics**: Available in results/study2\_cluster\_validation.csv **Full dendrogram**: Exported to results/study2\_hierarchical\_dendrogram.png

This methodology ensures that our thematic analysis represents genuine patterns in participant discourse rather than researcher-imposed interpretations, while maintaining statistical rigor through formal cluster validation procedures.