Study 2 Methodology Review: Implementation and Results

Erika Hernandez Acton

2025-06-05

# Study 2 Methodology Review Meeting

**Date:** June 5, 2025  
**Focus:** Study 2 qualitative analysis implementation and results

## Two Key Items for Review

### 1. **Hierarchical Clustering Methodology**

* Implemented data-driven approach using tidytext and hierarchical clustering
* Used Ward’s method for cluster analysis
* Results generated actual word clusters for thematic interpretation

### 2. **Speaker/Moderator Processing**

* Need to verify how moderator speech was handled in transcript processing
* Check if moderator utterances were filtered or included in analysis

## ✅ **STUDY 2 METHODOLOGY AND ACTUAL RESULTS**

### **Data Processing Implementation**

#### **📊 Processing Results**

* **Detection Rate:** 109/310 utterances (35.2%) contain SUD-specific terminology
* **Total SUD Tokens:** 2,253 tokens analyzed across 655 unique stems
* **Methodology:** Hierarchical clustering using Ward’s method (hclust(method = "ward.D2"))
* **Co-occurrence Analysis:** Used tidytext::pairwise\_count() for genuine term relationships

#### **🔬 Technical Implementation**

1. **Text Preprocessing** (study2\_text\_preprocessing.R):
   * Applied tidytext::unnest\_tokens() for word-level tokenization
   * Used 4 stopword sources including custom focus group terms
   * Applied Porter stemming via SnowballC::wordStem()
   * Detected SUD utterances using 53 substance-specific terms
2. **Co-occurrence Analysis** (study2\_cooccurrence\_analysis.R):
   * Used tidytext::pairwise\_count() for genuine co-occurrence relationships
   * Calculated word pair frequencies within SUD utterances
   * Created symmetric co-occurrence matrix for clustering
3. **Hierarchical Clustering**:
   * Applied Ward’s minimum variance method (hclust(method = "ward.D2"))
   * Used Euclidean distance on co-occurrence matrix
   * Cut dendrogram into 4 clusters (cutree(k = 4))

### **🎯 Data-Driven Cluster Results**

**CLUSTER 4 - PROFESSIONAL/FIELD FRAMEWORK (18.8% of SUD discourse)** *Most prominent theme with 357 total mentions* - **mental** (41), **helping** (33), **family** (30), **youre** (30), **abuse** (29) - **field** (29), **counselor** (27), **job** (25), **therapy** (24), **therapist** (23) - **ive** (23), **support** (23), **counseling** (20)

**CLUSTER 3 - SUBSTANCE/PERSONAL FRAME (11.9% of SUD discourse)** *226 total mentions* - **lot** (55), **dont** (54), **substance** (45), **person** (38), **health** (34)

**CLUSTER 1 - EMOTIONAL/AFFECTIVE DIMENSION (4.4% of SUD discourse)** *83 mentions* - **feel** (83) - emotional expressions and affective language

**CLUSTER 2 - RELATIONAL/SOCIAL DIMENSION (4.4% of SUD discourse)**  
*83 mentions* - **people** (83) - relational and social focus

### **📋 Next Steps**

1. **Review cluster word groupings** in results/study2\_cluster\_themes\_for\_naming.txt
2. **Assign thematic names** based on mathematically-derived clusters
3. **Investigate moderator speech processing** in preprocessing scripts
4. **Update manuscript methods** to reflect clustering approach

**Prepared by:** Erika Hernandez Acton  
**Purpose:** Study 2 methodology review and results documentation  
**Updated:** June 5, 2025