This is a preprint of the paper entitled

Approaches to Code Selection for Epistemic Networks

which has been published after peer-review in the proceedings for the International Conference on Quantitative Ethnography 2023

Please use the **citation** below to refer to this paper:

Árva, D., Jeney, A., Dunai, D., Major, D., Cseh, A., Zörgő, S. (2023). Approaches to Code Selection for Epistemic Networks. In: Arastoopour Irgens, G., Knight, S. (eds) Advances in Quantitative Ethnography. ICQE 2023. Communications in Computer and Information Science, vol 1895. Springer, Cham. https://doi.org/10.1007/978-3-031-47014-1_28

Approaches to Code Selection for Epistemic Networks

✉Dorottya Árva[1[0000-0003-3964-2708]], Anna Jeney[2[0000-0002-6037-9505]], Diána Dunai[3[0009-0002-8921-3434]], David Major[1[0000-0002-6108-9745]], Annamária Cseh[1[0009-0006-3664-7457]], Szilvia Zörgő[4[0000-0002-6916-2097]]

[1] Semmelweis University, 1089 Budapest, Hungary
[2] The Academy of Korean Studies, 13455 Seongnam, Republic of Korea
[3] Eötvös Loránd University, 1117 Budapest, Hungary
[4] Maastricht University, 6200 MD Maastricht, the Netherlands

arva.dorottya@med.semmelweis-univ.hu

Abstract. When employing unified, quantitative-qualitative methods such as Epistemic Network Analysis (ENA), the relative frequency of codes and their co-occurrence is of interest. However, in projects utilizing a large number of codes, if all codes are included, the interpretation of these models becomes challenging. In this paper, we provide three potential approaches to code selection. In the theory-based approach, code clustering and selection was founded on relevant literature or theory. In the insight-based approach, clusters of codes were defined by the grounded observations of researchers. Lastly, in the model-based approach, fully inclusive ENA models were generated to select codes for future models. We illustrated these approaches using data from our ongoing project that aims to measure the effects of a health education intervention on near-peer educators' understanding regarding the biopsychosocial model of health. All three approaches may be useful in guiding code selection for final ENA models or in providing a baseline for further refinement of model parameters. By outlining these approaches, this work contributes to discourse on making conscious and transparent decisions regarding ENA parameterization.

Keywords: Code Selection, Epistemic Network Analysis, Model Parameterization.

1 Introduction

1.1 Background

If aligned with ontological and epistemological assumptions, as well as research objectives, researchers may decide to transcribe and code qualitative data to identify patterns therein. Codes represent sets of concepts, gestures, expressions that capture relevant aspects of data (as defined by the research questions) and help researchers systematically categorize phenomena in their data [1, 2]. Provided a dataset has been coded systematically, the frequency, position, and interaction of codes can be subjected to further scrutiny, and quantitative models of the coded data can be generated [3]. Quantitative

models may inform qualitative insight and offer additional warrants to qualitative findings.

Models of quantified qualitative data can, for example, display the strength of association between codes, generally operationalized as co-occurrence frequencies. One way to model code co-occurrences is via Epistemic Network Analysis (ENA), which depicts the relative co-occurrence frequency of unique pairs of codes in designated segments of qualitative data [4, 5]. ENA models display two coordinated representations of the data in a two-dimensional space: (1) network graphs, where the nodes in the model correspond to the codes, and the edges represent the strength of association between codes, and (2) ENA scores, showing the relative position of each network as points. The position of nodes and the location of ENA scores in the constructed space are not arbitrary; network graphs can be used to interpret the meaning of ENA scores in terms of the network structures they represent [6]. For this reason, if a model parameter is altered, for example, by adding or removing a code, one alters the ENA model as well, which may have marked effects on its interpretation [7, 8].

ENA was originally designed to model a small set of codes [9] developed under the aegis of epistemic frames theory [10], but as it became applied to other theories in various fields, the scope and number of potential codes began to vary. Lefstein emphasizes the importance of "contextualization, performativity, co-construction, multi-modality and ideology in how we mean" and that these meaning-making processes constitute the foundation of hermeneutics [11]. He suggests these be most actively involved in the "precoding" or code development stage, but also subsequent to final coding, as the micro-analytic investigation of "select events" can not only validate our quantitative models, but also help discover novel topics and ideas that require further investigation [11]. Such iterative coding processes may generate a large number of codes, which, when placed into a single model, can present an overwhelming complexity.

The question of code selection has been addressed by Wang et al. as "parsimony": including the fewest number of codes that sufficiently explains the phenomenon of interest and retains interpretive alignment between the qualitative interpretation and the quantitative model [6]. The authors developed Parsimonious Removal with Interpretive Alignment (PRIA) to answer this challenge [6]. Yet, this technique entails having a 'gold standard model' to which more parsimonious (deflated) models are compared using statistical significance, goodness of fit, and interpretive alignment [6].

Especially if codes are developed inductively (grounded in or emerging from the data, as opposed to codes adopted from theory or a previous coding scheme), researchers may not have a clear 'gold standard model' prior to parameterization and analysis. Even with well-formulated research questions and goals, initial ENA models may serve solely exploratory purposes to identify salient themes and patterns within the data, which are subsequently examined in more detail [11]. In the following we discuss: what are some possible approaches to selecting codes to include in ENA models?

1.2 Epistemic Networks

Describing in detail how networks are generated does not fall within the scope of this paper (cf.: [1, 3, 5, 12]). Succinctly, in the process of wrangling, qualitative data is

segmented into lines (smallest meaningful pieces of data) and coded. Once in tabular form, ENA can process this coded data and produces a matrix with code co-occurrences, calculating the frequency of each unique pair of codes within given segments of data with a form of accumulation specified as a "stanza window". ENA aggregates the cumulative frequencies for each unit of analysis per "conversation" (a form of data segmentation); units are the totality of lines of data associated with a network within a model, and are usually defined as data providers or groups of data providers. The cumulative co-occurrence matrix for each unit is represented as a vector and forms an n-dimensional space. Vectors are normalized to account for varying amounts of data, which captures the relative frequency of code co-occurrences and also converts frequencies to fall between 0 and 1.

Subsequently, ENA applies a dimensional reduction procedure (singular value decomposition, SVD, or means rotation, MR) to reduce the n dimensions to just two. These two dimensions form the axes along which the unit vectors are then projected as points (ENA scores) into the two-dimensional space. The coordination of network graphs and plotted points means that the positions of the nodes can be used to interpret the dimensions forming the space and explain the positions of ENA scores. The x axis represents the dimension that explains the most variation in the co-occurrences, while the y axis represents the dimension that explains the most variance in the co-occurrences after the variance explained by the first dimension has been partialled out.

Thus, characteristics of the coded and segmented data, along with decisions in model parameterization, define epistemic networks and the space into which they are projected. Consequently, deciding on which codes to include in the model not only affects what relationships the networks display (i.e., which codes become nodes), but also determines the projection space and affects the interpretation of dimensions.

Precisely because ENA is sensitive to parameterization regarding codes, and because interpretive alignment with the qualitative data was paramount in its design as a visualization tool, we employ ENA to demonstrate approaches to selecting codes for co-occurrence frequency modeling and discuss potential implications for these choices.

1.3 Approaches to Selecting Codes

Theory-based. Some qualitative analytical frameworks that prescribe the researcher's stance to be as atheoretical as possible, such as Grounded Theory, where the aim of analysis is to generate a theory from the data [13]. Yet, most analytical procedures involve a dialectical relationship between theory and data, and advocate using theory to state preliminary assumptions and generate (sets of) codes. ENA can, in turn, be employed to explore assumptions about the relationships among codes [6]. Relevant literature and theory are most frequently employed at research design and code development, but can also scaffold methodological choices in modeling.

Insight-based. Once data has been collected (or even during data collection itself), researchers often engage with their data and gain qualitative insights. These observations (e.g., constructs of interest, perceived patterns, inconsistencies or atypical examples)

may contribute to code development, especially if codes are created inductively. As researchers develop and test the applicability of their codes, they may formulate "favorite theories" about their data [11], a grounded understanding leading to preconceptions about critical relationships (or their absence) among certain codes [14]. A more mature set of observations (e.g., based on initial coding or hermeneutic analysis) may be referred to as a theme: a constellation of codes the researcher identifies as meaningful and significant [15]. Such grounded assumptions, "favorite theories", or themes may serve as the basis for selecting codes to include in an ENA model.

Model-based. Provided the use of theory was not justified or possible, and qualitative engagement with the data was not warranted or did not yield any observations (or yielded too many), another, more quantitative approach to selecting codes may be appropriate. Since epistemic networks are projected into a meaningful space, the position of nodes can be employed to formulate assumptions about relationships among codes. All codes can be included in a single model to inspect this space, albeit this may create a highly dense network, and nodes may even eclipse each other. Yet, the clustering of certain codes, or code positions relative to the axes (dimensions) creating this space, may offer insight into how codes relate to each other within the entire dataset. These insights can then be mobilized to create subsets of codes and their respective models.

In the following, we use data from our ongoing project to elaborate examples for all three suggested approaches and to discuss their implications regarding model construction and interpretation. First, we introduce the context of our research, our goals, employed methods, and disclose our codes. Next, each example will reflect a viable means of selecting codes for model construction and a brief discussion of the generated results. Subsequently, we discuss the implications of these approaches.

2 Data in Use

2.1 Project Overview

Several models of health and illness share the understanding that health is determined by a number of factors and their interactions [16, 17]. Bircher [18] states that health emerges from interactions among individual, social, and environmental factors. A widely known model capturing the interplay of such factors is referred to as the *biopsychosocial model* of health [16]. Effective health promotion, prevention, and health care leverage this model, and congruently, so do successful health education programs.

The Balassagyarmat Health Education Program (BEP) was a school-based health education intervention, run between 2018-2021, aiming to improve the health behavior of high school students in Balassagyarmat, a city in a socioeconomically disadvantaged region in Hungary [19]. Interactive offline and online sessions were designed using gamification and peer education, a commonly employed method in school health education [20]. Students of medicine, as near-peer educators, taught high schoolers for a year. The program, developed by a multidisciplinary team at Semmelweis University,

covered a wide range of health-related topics in nine modules: healthy nutrition and physical activity, smoking, alcohol, drugs, reproductive and mental health, infection control, and basic life support. Educators received 18 hours of training each semester, which focused on the material they delivered to high school students and the biopsychosocial model of health.

Upon completion of the BEP intervention, we not only wanted to explore the effects it had on high school students as the primary target group, but also on the educators themselves. To achieve the latter, we are currently comparing the educators' understanding of the biopsychosocial model to those of medical students' who did not participate in the intervention. We assumed there is a correlation between exposure to the intervention and knowledge on the biopsychosocial model of health.

3 Methods

Both subsamples were recruited from Semmelweis University, Budapest, Hungary: 1) BEP educators (who learnt all modules and taught in-person; n=9), 2) controls (medical students pair-matched for academic year and sex; n=9). We conducted simulation interviews (a form of knowledge elicitation, cf.: [21]) where cognitive task analysis [22] was performed on visual stimuli. Interviewees were probed via a standardized protocol on declarative and procedural knowledge on the determinants of health and their interplay. We also administered a survey to collect sociodemographic data and self-reported health behavior. Data were collected online by pairs of seven trained interviewers between December 2021 and February 2023; interviews lasted 99 minutes on average.

Codes were developed in several stages in a guided inductive process based on the eight[1] modules of the intervention constituting parent codes. For a more detailed description of code development, see our preregistration (https://osf.io/hjs5b). The final codebook contained two code clusters: substantive codes reflecting the intervention modules and "metacodes" capturing aspects of our data that spanned across substantive codes. Substantive codes were hierarchical, comprising two levels of abstraction; metacodes were clustered in a flat structure. Table 1 contains the simplified version of our final codebook; the more detailed version is available online: https://osf.io/t8xh5.

Table 1. Simplified version of our codebook.

Parent code	Child code	Code definition
Nutrition	Unhealthy nutrition	Malnutrition, bad eating habits, bad food choices
	Healthy nutrition	Healthy eating habits, good food choices
	Unhealthy weight	Overweight or underweight, energy imbalance
	Healthy weight	Healthy weight, energy balance

[1] We decided not to include the topic of basic life support in code development because it pertained to health achieved by proxy in emergency situations.

Physical activity	Healthy exercise	Healthy quality and quantity of exercise
	Unhealthy exercise	Inactivity or too much exercise
	Adequate sleep	Right hours and quality of sleep, no disturbances
	Inadequate sleep	Too little/much sleep; bad quality, disturbances
Smoking	Active smoking	Using tobacco or nicotine products
	Passive smoking	Exposure to someone else's smoking
Alcohol	Alcohol unhealthy	Dysfunctional, chronic, uncontrolled drinking
	Alcohol acceptable	Moderate, controlled, or occasional drinking
Drugs	Drugs unhealthy	Unhealthy and serious effects of drugs
	Drugs acceptable	Less harmful drugs and experimentation
Sex	Healthy sex	Physical and mental health promoting sex life
	Unhealthy sex	Physical or mental health harms of sex life
Mental health	Social support	Good relationships, positive social influences
	Social negative	Bad relationships, negative social influences
	Mental well-being	Personal mental health and self-understanding
	Mental ill-being	Mental health problems, lack of self-acceptance
Infection control	Hygiene	Basic hygiene and cleanliness
	Lack of hygiene	Lack of hygiene and cleanliness
Metacodes	Addiction	Addiction to any substance or behavior
	Abstinence	Refraining from exhibiting a certain behavior
	Finance	Money as a factor in health or ill-health
	Regulations	Laws and regulations on alcohol and drugs
	Access	In/availability of services and products
	Preventive health care	Non/use of healthcare for preventive purposes
	Adherence	Use of medication according to prescription
	Ability	Knowledge, skills and responsibility for health
	Physical environ-ment	Health effects of housing conditions, physical en-vironment, geographical location

Interviews were transcribed verbatim and anonymized. Sentences comprised our lowest level of segmentation; we employed the Reproducible Open Coding Kit (ROCK) R

package[2] to place sentences on separate lines and designate a unique utterance identifier to each. We used the Interface for the Reproducible Open Coding Kit (iROCK)[3] to code and segment our data. Five researchers in our team "specialized" in a set of codes each (5-7 codes from the total 31), and one researcher was responsible for segmenting transcripts according to visual stimuli (three pictures used during the simulation interviews) and health determinants (elaboration of determinants by interviewees).

Coding was performed on the level of sentences. The six coded versions of each interview were merged with the ROCK R package and exported into tabular format where sentences comprised rows, and columns were designated for each of our codes and types of segmentation. If a code was present in a line of data, it was indicated with a 1, the absence of a code with a 0. Stimuli-based segmentation was categorical, determinant-based segmentation received ordinal numbering within each interview. The interview protocol, visual stimuli, comprehensive codebook, coded dataset, and other materials are available in our repository: https://osf.io/ynjv4.

In the following, we explore three approaches to selecting codes for our ENA models. Our coding is currently in progress, hence we utilize only the intervention group data for our examples. Code names and quotes (translated by the authors), marked with the case ID [cid] of participants, are in italic.

4 Results

4.1 Selecting Codes Based on Theory

In the theory-based approach, we selected codes founded on relevant literature and theory. Our intervention aimed to improve the health status of adolescents, hence theory was aligned with this target group. Substance use behaviors often begin during adolescence [23] and are associated with both short-, and long-term health problems [24]. The smoking, alcohol, and drug use-related disease burden is significant in the European Union, especially in Hungary [25], therefore the prevention of adolescent substance use and addictions are relevant targets for public health interventions. Planning such preventive measures can build on the growing literature of *risk and protective factors* for substance use. Numerous risk factors have been identified, such as 1) substance use among friends, peers, and family, 2) high perceived accessibility of drugs, and 3) mental health problems, such as depressive symptoms and anxiety [26, 27]. Protective factors have also been identified, for example: 1) psychosocial competencies (e.g., assertiveness, coping, mindfulness, and optimism), 2) secure attachment and family cohesion, 3) successful integration in school, and 4) anti-substance policies [26–29]. BEP covered both risk and protective factors in its modules on alcohol, smoking, and drugs [19]. To explore the associations between substance use and its risk and protective factors in BEP educator narratives, we generated an ENA model by selecting all children of parent codes: *Smoking*, *Alcohol*, *Drugs*, *Mental health* as well as metacodes *Ability*, *Addiction*, *Abstinence*, *Access*, and *Regulation*.

[2] https://rock.opens.science
[3] https://i.rock.science

Table 2 contains ENA model parameterization. Units were specified in a nested structure which allowed us to create a mean network for the intervention subsample (Group), participants (cid), and stimuli. Conversation was defined as determinants, which were subsections within narratives on specific stimuli. Code co-occurrences were accumulated with a weighted whole conversation stanza window, which was justified by theory regarding the meaning of co-occurrences in the text and by iteration between qualitative insights and quantitative models.

Table 2. Parameters of the Epistemic Network Analysis (ENA) model generated for illustrating the theory-based approach to selecting codes.

Unit	Group>cid>stimuli
Conversation	Determinants
Stanza window	Weighted whole conversation
Codes	Active smoking, Passive smoking, Alcohol unhealthy, Alcohol acceptable, Drugs unhealthy, Drugs acceptable, Social support, Social negative, Mental well-being, Mental ill-being, Addiction, Abstinence, Finance, Regulations, Access, Ability
Minimum edge weight	0.04
Projection	SVD1 (12.7%); SVD2 (11.6%)

Fig. 1 displays the network reflecting participants' understanding of risk and protective factors regarding substance use. In terms of risk factors, there was a strong emphasis on substance use in social situations, as reflected in the connection between *Social support* and *Active smoking*: *"These gatherings are essential to build relationships among teenagers, and alcohol is part of these, even if smoking should not be"* (cid 101). Peer-pressure, manifesting in the connection between *Social negative* and *Active smoking* and *Drugs unhealthy*, also appeared as a related risk factor. This signified that peers and culture have a powerful negative influence on an individual's substance use: *"Well, unfortunately we're susceptible to influence and we like to be alike, and that includes if it's drug use or even just smoking"* (cid 101). Risk associated with mental health issues was also present in our network via its connection to dangerous drug use (*Drugs unhealthy*), reflecting the view that psychological distress may lead to drug use as a coping mechanism and drug use may result in mental health issues. High perceived accessibility of substances, however, was not present in the narratives as a risk factor: no connections were exhibited between the substance use codes and *Access* or *Regulations.*

Considering protective factors, *Mental well-being* and *Regulations* did not have connections with substance use codes, but the role of assertiveness appeared in the connection between *Ability* and *Addiction: "[It is healthy] if someone is aware of the addictiveness of substances and of his/her proneness to addictions* […] *and can draw a line and keep it, when facing these things"* (cid 109). Thus, participants related substance use to both risk and protective factors, however, representation of the latter was less pronounced.

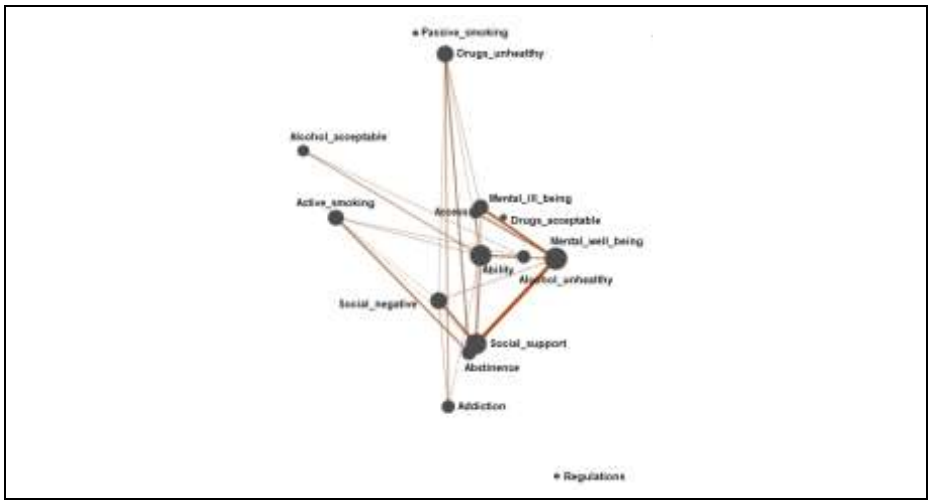

Fig. 1. Mean epistemic network of the educators of the Balassagyarmat Health Education Program (intervention group) on risk and protective factors of substance use, illustrating the theory-based approach to selecting codes. Codes are represented by nodes (circles); node size and edge thickness indicate the relative frequency of code co-occurrence.

4.2 Selecting Codes Based on Insight

Employing the insight-based approach to code selection, we leveraged the grounded assumptions of researchers from the stages of data collection, codebook development, and final, deductive coding. One observation was that, according to participants, healthy nutrition was dependent on the availability of healthy food in shops (e.g., whole grain products), and the latter was said to be determined by the size of the municipality of residence and the financial situation of individuals. To systematically explore this theme, we constructed a model including children of the parent code *Nutrition* and meta-codes *Access, Finance* and *Physical environment*. Table 3 summarizes ENA model parameterization. The stanza window designation was changed compared to the previous model to optimize the model's ability to fit with qualitative insights.

Table 3. Parameters of the Epistemic Network Analysis (ENA) model generated for illustrating the insight-based approach to selecting codes.

Unit	Group>cid
Conversation	Determinants
Stanza window	Moving stanza of 4 lines
Codes	Unhealthy nutrition, Healthy nutrition, Unhealthy weight, Healthy weight, Access, Finance, Physical environment
Minimum edge weight	0.04
Projection	SVD1 (45.8%); SVD2 (25.7%)

The network in Fig. 2 displays our validated preliminary insight on nutrition. *Access* displayed connections with both *Healthy nutrition* and *Unhealthy nutrition*. The connections manifested as nutritional habits being highly dependent on the availability of products that can either serve health (e.g., home grown/farmed products) or be detrimental to it (e.g., fast food). *Physical environment* exhibited connections to the nutrition codes as well, meaning that some healthy nutritional products are more commonly consumed in the countryside, as the abovementioned home grown products. However, this connection also encompassed the hazards of food processing at home (e.g., smoking meat can result in increased carcinogen content). *Physical environment* displayed a strong connection with *Access* and *Finance*, signifying that place of residence determines the availability of certain products and services (e.g., sewage system, home-grown products, education) and that the financial status of people influences health. An exemplar of these co-occurrences is captured by a participant as follows: *"We can suggest [people living in the countryside] to buy processed meat of a better quality [because they are healthier], but those are way more expensive* […] *Vegetables are expensive too.* […] *Fortunately, those living in the countryside can grow some of their own"* (cid 108). Yet, in contrast with our preliminary insight, it was *Physical environment*, not *Access*, that showed the most marked connections with nutrition codes.

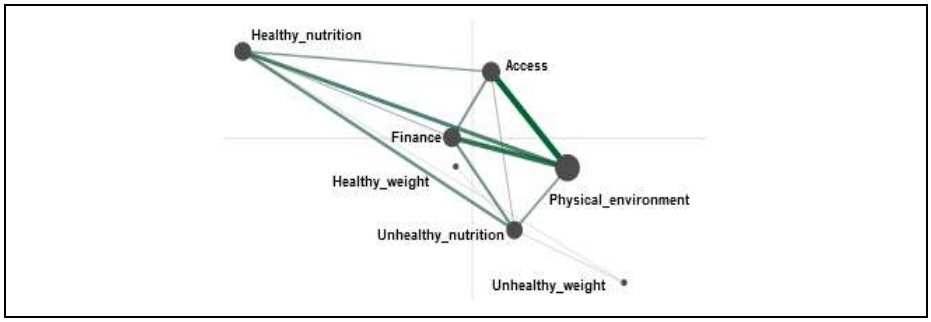

Fig. 2. Mean epistemic network of the educators of the Balassagyarmat Health Education Program (intervention group) on determinants of healthy nutrition, illustrating the insight-based approach to selecting codes. Codes are represented by nodes (circles); node size and edge thickness indicate the relative frequency of code co-occurrence.

4.3 Selecting Codes Based on a Full Model

When applying the model-based approach, we generated an ENA model with all codes, observed the ENA projection space, and made selection decisions for future models based on node positioning. Table 4 contains the parameters of this ENA model. The stanza window designation in this model was chosen arbitrarily, as no theory or qualitative insight could guide this decision.

Table 4. Parameters of the Epistemic Network Analysis (ENA) model generated for illustrating the model-based approach to selecting codes.

Unit	Group>cid>factors
Conversation	Determinants
Stanza window	Moving stanza of 4 lines
Codes	All 31
Minimum edge weight	0.00
Projection	SVD1 (6.1%); SVD2 (4.5%)

The x axis, explaining the most variance in our data, was constituted by *Access, Physical environment,* and *Finance* on the one hand, and *Ability, Mental well-being,* and *Social support* on the other. This dimension, shown in Figure 3, can be interpreted as contrasting environmental and socioeconomic determinants with individual ones; the former characterized by lower individual control and greater determination by the broad societal and environmental systems, the latter by personal and interpersonal aspects of health over which the individual may have more control. Defining this dimension could thus help form a grounded theory on how code clusters can be created; the two poles of the dimension and associated codes are potentially separate models (e.g., an ENA model displaying individual factors in health). Congruently, the y axis can be used to enrich the theory built from the first dimension or to construct a novel grounded theory.

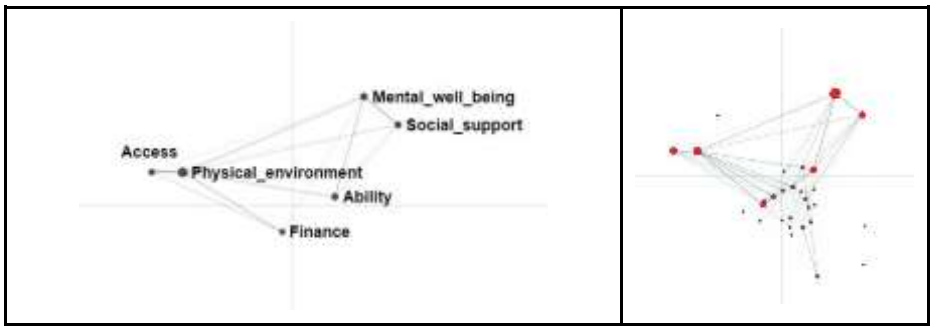

Fig. 3. *Right:* A full epistemic network of the educators of Balassagyarmat Health Education Program (intervention group) on the determinants of health. Nodes highlighted in red are represented in the left plot. *Left:* Specific codes that drive the x axis. Codes are represented by nodes (circles); node size and edge thickness indicate the relative frequency of code co-occurrence.

Viewing the ENA projection space in terms of quadrants (upper right, upper left, lower right, lower left) may also inform code selection. Codes taking on a proximal position in the space exhibit similar connections to all other codes in the dataset. By this logic, using quadrants (or node placement in general) can be employed to define code clusters based on co-occurrence patterns. Depending on analytical goals, we may want to investigate connections among codes that exhibit similar co-occurrence patterns in the data or codes that differ in how they connect to other codes.

If we wanted to delve deeper into connections among codes that exhibit a similar co-occurrence pattern in the dataset, we could select codes that are proximal in the ENA space and generate a separate model for them. For example, as shown in Fig. 4, we may choose codes *Mental ill-being, Active smoking, Ability, Social negative, Alcohol unhealthy,* and *Drugs unhealthy* to examine their interactions more closely (purple cluster, Fig. 4, left). A separate model could be created for codes *Healthy nutrition, Unhealthy nutrition, Unhealthy exercise,* and *Passive smoking* for the same purpose (teal cluster, Fig. 4, left). This would contribute to conclusions such as individual decisions are key to substance use behaviors, and peer pressure and mental health issues make healthy choices more difficult.

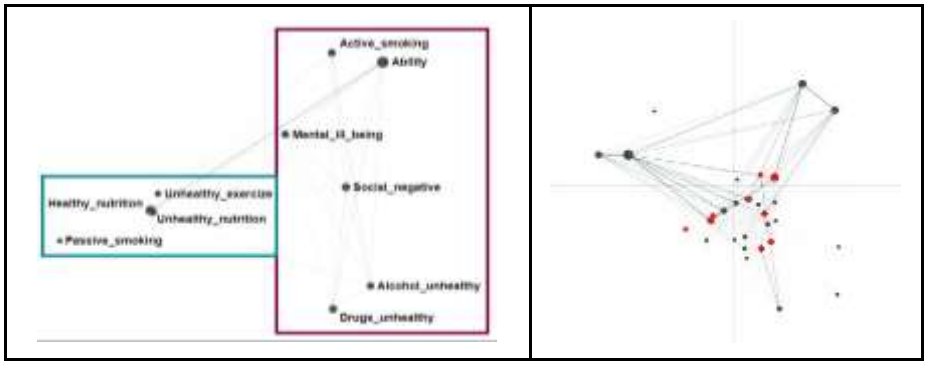

Fig. 4. *Right:* A full epistemic network of the educators of the Balassagyarmat Health Education Program (intervention group) on the determinants of health. Nodes highlighted in red are represented in the left plot. *Left:* Two clusters of codes that exhibit a similar co-occurrence pattern in the dataset. Codes are represented by nodes (circles); node size and edge thickness indicate the relative frequency of code co-occurrence.

Conversely, if we wanted to create a model that elaborates the connection between clusters of codes with differing co-occurrence patterns, we could create groups of proximally situated codes and dichotomize their occurrence into a novel, derived variable; in essence, a new parent code would be created. This would entail designating a new column in the dataset that is coded line-by-line as other code columns.

Code proximity, as the basis for derived parent codes, can be inspected in the default ENA space or with an alternative plot called a unit circle[4] shown in Fig. 5. Codes can be grouped in several ways, for example, based on quadrant or proximity to an axis (e.g., *Hygiene, Lack of Hygiene, Regulations* and *Alcohol acceptable* in Fig. 5).

[4] The unit circle node layout positions the network nodes where a line drawn from the origin to each node in the default ENA layout intersects a unit circle. This places the nodes such that they are all equidistant from the origin.

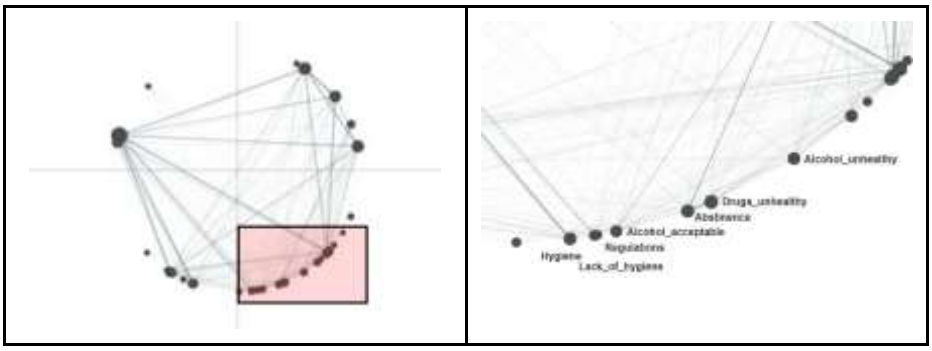

Fig. 5. *Left:* Unit circle node layout of a full epistemic network of the educators of the Balassagyarmat Health Education Program (intervention group) on the determinants of health. *Right:* A closer view of its highlighted area in the right plot. Codes are represented by nodes (circles); node size and edge thickness indicate the relative frequency of code co-occurrence.

5 Discussion

In this paper, we aimed to demonstrate viable approaches to selecting codes for ENA models. The theory-based approach entailed using literature or theory to formulate assumptions and code clusters that affirm or challenge those. The insight-based approach relied on the researchers' grounded observations as the basis of code selection. Lastly, in the model-based approach, an ENA projection space was created that included all possible codes; selection relied on inspecting the position of codes in this space.

Theory is an integral and inherent element of research [30]. Yet, guided inductive code development (basing codes/codebook on theory or literature) or deductive code development (adopting codes/codebook from others) is not without challenges. Often, codebooks may not be made public by researchers or may not reach the level of specificity needed to adopt and apply codes reliably [31]. Furthermore, constructs may differ in how they are defined even among authors using the same theoretical framework. Thus, use of identical construct labels does not necessarily imply that they are measuring the same phenomenon [32, 33].

The effects of researcher biases and preferences are ubiquitous throughout research, and are even embedded in analytical tools [34, 35]. One might argue that constructing a model of the data that reflects qualitative insights only generates findings prone to the confirmation bias [36][5]. Preregistering research and employing credibility strategies, such as reflexivity, iterative codebook construction, and social moderation during the code development phase, or respondent validation and peer debriefing during the analysis phase, may prove to be effective in reducing researcher bias [37]. The coded dataset should already be a scrutinizable output of transparent and systematic processes. Thus, basing code selection on insights gained during various phases of research does not

[5] We are intentionally sidestepping the crucial epistemological and ontological question of whether biases should be minimized in scientific outputs or employed as an analytical tool.

necessarily increase the effects of bias, in fact, the iteration between qualitative understanding and quantitative model may even disprove initial researcher assumptions.

When working with many codes, the level of difficulty in interpreting pairwise connections in an epistemic network rises, and nodes may start to eclipse each other, lending challenges to visual inspection [9]. A model-based approach to code selection may be appropriate if, for example, no theory is employed (because, e.g., generating a grounded theory is an analytical objective [13]) or no qualitative insights can reliably be gained (because, e.g., insights of interest depend on multiple attributes of data providers). Albeit, one should exercise caution in solely selecting codes that exhibit the strongest connections and excluding those with weak or no connections from future models, as some crucial connections may be hidden behind the effects of "dominant codes" [9] and weak connections may play a significant role in explaining variance. Furthermore, code pairs exhibiting no connection may be a critical aspect of the findings as well.

The most notable limitation of this study is that in order to elaborate the three approaches to code selection, we did not address other questions in model parameterization, and did not discuss how those decisions affect networks and their interpretation. Choices pertaining to co-occurrence accumulation and aggregation (e.g., operationalization of unit, conversation, and stanza window) are especially crucial [1, 14]. A further limitation is that the specification of our units was primarily guided by the number of codes we wanted to include in our model. Aiming to keep the number of units higher than the number of codes to ensure statistical validity [12] resulted in varying unit designations across the demonstrated approaches.

6 Conclusions

The presented approaches to code selection offer strategies for addressing a large number of codes in analysis, both to uncover fine details and to outline broader associations. The networks that are produced in any of these approaches can be considered initial, exploratory models aiding a dialectic with the researcher's qualitative understanding of the data, or can constitute the final models included in the write-up of results. Our paper aims to spark further discussion on modeling interaction among many codes, a common challenge in conveying qualitative research findings.

7 Acknowledgements

We wish to thank our interviewees for their participation and sacrificing their time. This study was supported by the ÚNKP-22-3-I-SE-11 New National Excellence Program of the Ministry for Culture and Innovation from the Source of the National Research, Development and Innovation Fund and by the European Union and the Hungarian State (grant number: EFOP-3.4.3-16-2016-00007). This project also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 101028644, as well as from University Fund

Limburg/SWOL. The opinions, findings, and conclusions do not reflect the views of the funding agency, cooperating institutions, or other individuals.

References

1. Zörgő S (2023) Segmentation and Code Co-occurrence Accumulation: Operationalizing Relational Context with Stanza Windows. In: Damşa C, Barany A (eds) Advances in Quantitative Ethnography. Springer Nature Switzerland, Cham, pp 146–162
2. Cai Z, Siebert-Evenstone A, Eagan B, et al (2019) nCoder+: A Semantic Tool for Improving Recall of nCoder Coding. In: Eagan B, Misfeldt M, Siebert-Evenstone A (eds) Advances in Quantitative Ethnography: First International Conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, Proceedings. International Society for Quantitative Ethnography, pp 41–54
3. Shaffer D (2017) Quantitative Ethnography. Cathcart Press
4. Shaffer DW, Collier W, Ruis AR (2016) A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. J Learn Anal 3:9–45
5. Zörgő S, Peters G-J (2023) Using the Reproducible Open Coding Kit & Epistemic Network Analysis to model qualitative data. Health Psychol Behav Med 11:. https://doi.org/10.1080/21642850.2022.2119144
6. Wang Y, Swiecki Z, Ruis A, Shaffer DW (2021) Simplification of Epistemic Networks Using Parsimonious Removal with Interpretive Alignment. In: Ruis AR, Lee SB (eds) Advances in Quantitative Ethnography: Second International Conference, ICQE 2020, Malibu, CA, USA, February 1-3, 2021, Proceedings. Springer, pp 137–151
7. Siebert-Evenstone AL, Arastoopour G, Collier W, et al (2017) In Search of Conversational Grain Size: Modelling Semantic Structure Using Moving Stanza Windows. J Learn Anal 4:123–139
8. Zörgő S, Swiecki Z, Ruis AR (2021) Exploring the Effects of Segmentation on Semi-Structured Interview Data with Epistemic Network Analysis. In: Advances in Quantitative Ethnography. Communications in Computer and Information Science Series, Eds. Ruis AR and Lee, SB. Springer Nature, Switzerland, pp 78–90
9. Mello RF, Gašević D (2019) What is the Effect of a Dominant Code in an Epistemic Network Analysis? In: Eagan B, Misfeldt M, Siebert-Evenstone A (eds) Advances in Quantitative Ethnography: First International Conference, ICQE 2019, Madison, WI, USA, October 20–22, 2019, Proceedings. Springer, pp 66–76
10. Shaffer DW (2006) Epistemic frames for epistemic games. Comput Educ 46:223–234
11. Lefstein A (2022) Interpretation in Linguistic Ethnography: Some Comments for Quantitative Ethnographers. Work Pap Urban Lang Literacies 297:
12. Bowman D, Swiecki Z, Zhiqiang C, et al (2021) The Mathematical Foundations of Epistemic Network Analysis. In: Advances in Quantitative Ethnography. Communications in Computer and Information Science Series., Eds. Ruis AR and Lee, SB. Springer Nature, Switzerland, pp 91–105
13. Glaser BG, Strauss AL (1967) The discovery of grounded theory: strategies for qualitative research. Aldine Publishing, Chicago

14. Zörgő S, Brohinsky J (2023) Parsing the Continuum: Manual Segmentation of Monologic Data. In: Damşa C, Barany A (eds) Advances in Quantitative Ethnography. Springer Nature Switzerland, Cham, pp 163–181
15. Braun V, Clarke V (2012) Thematic analysis. American Psychological Association
16. Engel GL (1977) The Need for a New Medical Model: A Challenge for Biomedicine. Science 196:129–136. https://doi.org/10.1126/science.847460
17. Dahlgren G, Whitehead M (1991) Policies and strategies to promote social equity in health. Background document to WHO - Strategy paper for Europe. Inst Futur Stud Arbetsrapport 14:
18. Bircher J, Kuruvilla S (2014) Defining health by addressing individual, social, and environmental determinants: New opportunities for health care and public health. J Public Health Policy 35:363–386. https://doi.org/10.1057/jphp.2014.19
19. Eörsi D, Árva D, Herzeg V, Terebessy A (2020) Komplex iskolai egészségfejlesztő program a COM-B modell tükrében [Introduction to a complex school-based health education program from the COM-B model's perspective]. Egészségfejlesztés 61:36–47. https://doi.org/10.24365/ef.v61i1.540
20. Mellanby AR, Rees JB, Tripp JH (2000) Peer-led and adult-led school health education: a critical review of available comparative research. Health Educ Res 15:533–545. https://doi.org/10.1093/her/15.5.533
21. Cooke NJ (1994) Varieties of knowledge elicitation techniques. Int J Hum-Comput Stud 41:801–849. https://doi.org/10.1006/ijhc.1994.1083
22. Crandall B, Klein G, Hoffman R (2006) Working Minds: A Practitioner's Guide to Cognitive Task Analysis
23. Inchley J, Currie D, Budisavljević S, et al (2020) FINDINGS FROM THE 2017/2018 HEALTH BEHAVIOUR IN SCHOOL-AGED CHILDREN (HBSC) SURVEY IN EUROPE AND CANADA INTERNATIONAL REPORT VOLUME 1. KEY FINDINGS Spotlight on adolescent health and well-being Spotlight on adolescent health and well-being
24. Lee H, Henry KL (2022) Adolescent Substance Use Prevention: Long-Term Benefits of School Engagement. J Sch Health 92:337–344. https://doi.org/10.1111/josh.13133
25. OECD, Policies EO on HS and (2021) Hungary: Country Health Profile 2021
26. Bozzini AB, Bauer A, Maruyama J, et al (2021) Factors associated with risk behaviors in adolescence: a systematic review. Rev Bras Psiquiatr Sao Paulo Braz 1999 43:210–221. https://doi.org/10.1590/1516-4446-2019-0835
27. Nawi AM, Ismail R, Ibrahim F, et al (2021) Risk and protective factors of drug abuse among adolescents: a systematic review. BMC Public Health 21:2088. https://doi.org/10.1186/s12889-021-11906-2
28. Das JK, Salam RA, Arshad A, et al (2016) Interventions for Adolescent Substance Abuse: An Overview of Systematic Reviews. J Adolesc Health 59:S61–S75. https://doi.org/10.1016/j.jadohealth.2016.06.021
29. Evren C, Dalbudak E, Evren B, Demirci AC (2014) High risk of Internet addiction and its relationship with lifetime substance use, psychological and behavioral problems among 10(th) grade adolescents. Psychiatr Danub 26:330–339
30. Collins CS, Stockton CM (2018) The Central Role of Theory in Qualitative Research. Int J Qual Methods 17:1609406918797475. https://doi.org/10.1177/1609406918797475

31. Zörgő S, Peters G-JY, Porter C, et al (2022) Methodology in the Mirror: A Living, Systematic Review of Works in Quantitative Ethnography. In: Wasson B, Zörgő S (eds) Advances in Quantitative Ethnography: Third International Conference, ICQE 2021, Virtual Event, November 6-11, Proceedings. Springer, pp 144–159
32. Peters G-JY, Crutzen R (2017) Pragmatic nihilism: How a Theory of Nothing can help health psychology progress. Health Psychol Rev 11:103–121. https://doi.org/10.1080/17437199.2017.1284015
33. West R, Godinho CA, Bohlen LC, et al (2019) Development of a formal system for representing behaviour-change theories. Nat Hum Behav 3:526–536. https://doi.org/10.1038/s41562-019-0561-2
34. Arastoopour Irgens G, Eagan B (2023) The Foundations and Fundamentals of Quantitative Ethnography. In: Damşa C, Barany A (eds) Advances in Quantitative Ethnography: Fourth International Conference, ICQE 2022, Copenhagen, Denmark, October 15–19, 2022, Proceedings. Springer, pp 3–16
35. Vaandering D, Reimer KE (2021) Relational Critical discourse analysis: a methodology to challenge researcher assumptions. Int J Qual Methods 20:16094069211020904
36. Nickerson R (1998) Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Rev Gen Psychol 2:175–220
37. Zörgő S (2021) Preregistration Template for Qualitative and Quantitative Ethnographic Studies. https://doi.org/10.17605/OSF.IO/TGK49