

Psychological Methods

A Computational Method to Reveal Psychological Constructs From Text Data

Alina Herderich, Heribert H. Freudenthaler, and David Garcia

Online First Publication, September 19, 2024. <https://dx.doi.org/10.1037/met0000700>

CITATION

Herderich, A., Freudenthaler, H. H., & Garcia, D. (2024). A computational method to reveal psychological constructs from text data.. *Psychological Methods*. Advance online publication. <https://dx.doi.org/10.1037/met0000700>

A Computational Method to Reveal Psychological Constructs From Text Data

Alina Herderich¹, Heribert H. Freudenthaler², and David Garcia^{1, 3, 4}

¹Department of Computer Science and Biomedical Engineering, Graz University of Technology

²Institute of Psychology, University of Graz

³Department of Politics and Public Administration, University of Konstanz

⁴Complexity Science Hub Vienna


Abstract


When starting to formalize psychological constructs, researchers traditionally rely on two distinct approaches: the quantitative approach, which defines constructs as part of a testable theory based on prior research and domain knowledge often deploying self-report questionnaires, or the qualitative approach, which gathers data mostly in the form of text and bases construct definitions on exploratory analyses. Quantitative research might lead to an incomplete understanding of the construct, while qualitative research is limited due to challenges in the systematic data processing, especially at large scale. We present a new computational method that combines the comprehensiveness of qualitative research and the scalability of quantitative analyses to define psychological constructs from semistructured text data. Based on structured questions, participants are prompted to generate sentences reflecting instances of the construct of interest. We apply computational methods to calculate embeddings as numerical representations of the sentences, which we then run through a clustering algorithm to arrive at groupings of sentences as psychologically relevant classes. The method includes steps for the measurement and correction of bias introduced by the data generation, and the assessment of cluster validity according to human judgment. We demonstrate the applicability of our method on an example from emotion regulation. Based on short descriptions of emotion regulation attempts collected through an open-ended situational judgment test, we use our method to derive classes of emotion regulation strategies. Our approach shows how machine learning and psychology can be combined to provide new perspectives on the conceptualization of psychological processes.

Translational Abstract

Psychologists often study behavioral and cognitive concepts, which are abstract phenomena not directly observable. To investigate these concepts, such as emotions, researchers need clear definitions and indicators that signal their presence or absence. Past research has relied on two main approaches: expert knowledge and self-report questionnaires, or in-depth analysis of text often from a small set of interviews. Both approaches have limitations, either due to specific ideas that researchers have about a concept or due to restricted data processing capabilities. In this paper, we present a novel method for defining psychological concepts that unites the detail of text analysis with the ability to process large datasets efficiently. Using natural language processing techniques, we convert open-ended survey responses into numerical values. Through clustering, we identify classes of the concept of interest. Moreover, we introduce measures to assess the impact of data collection on the results and to assess the meaningfulness of classes according to human

Douglas Steinley served as action editor.

Alina Herderich  <https://orcid.org/0000-0002-2940-600X>

David Garcia  <https://orcid.org/0000-0002-2820-9151>

An early conceptualization of this research was presented at the CERE 2022, the Falling Walls Lab Austria 2022, and the European Forum Alpbach 2022. The work was further presented at SPSP 2024 and SAS 2024. All the listed conferences are non-archival. The project is introduced on the first authors website: <https://hai-lina.github.io/>. All code necessary for the replication of the emotion regulation example is made publicly available on GitHub (<https://github.com/Hai-Lina/construct-mining-pipeline>). Emotion regulation strategy descriptions, as well as data from the intrusion task, are made publicly available on GitHub (<https://github.com/Hai-Lina/construct-mining-pipeline>). A preprint of this article can be found at: <https://psyarxiv.com/s64tm> (<https://10.31234/osf.io/s64tm>). The current research was approved by the Institutional Review Board (IRB) of the University of Graz under application number GZ. 39/122/63 ex 2021/22. The authors thank Jana Lasser for her critical input during the entire research and writing process. The authors also thank Segun Aroyehun, Indira Sen, and

Mirta Galesic for providing their feedback on a draft version of this article. The authors declare no competing interests. The authors refrain from publishing personally identifiable information to preserve the anonymity of participants. The authors used ChatGPT 3.5 to rephrase and refine a human-written version of the translational abstract. The authors did not use large language models for any other task or part of the article.

Alina Herderich served as lead for conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, visualization, writing—original draft, writing—review and editing. Heribert H. Freudenthaler contributed equally to resources and served in a supporting role for conceptualization, data curation, methodology, supervision, and writing—review and editing. David Garcia served as lead for funding acquisition and supervision, contributed equally to conceptualization, methodology, visualization, and writing—review and editing, and served in a supporting role for formal analysis and software.

Correspondence concerning this article should be addressed to Alina Herderich, Department of Computer Science and Biomedical Engineering, Graz University of Technology, Sandgasse 36/III, 8010 Graz, Austria. Email: alina.herderich@tugraz.at

judgment. To demonstrate the effectiveness of our method, we develop a classification system for emotion regulation strategies using emotion regulation descriptions collected from an open-ended survey. Our method showcases the potential of machine learning to enhance psychological research and refine the conceptualization of psychological phenomena.

Keywords: constructs, psychological assessment, natural language processing, clustering, emotion regulation

Arriving at precise and comprehensive definitions of psychological constructs remains a challenge. This difficulty stems in large part from the inherent interest of psychological science in latent concepts that are not directly observable. For example, when developing questionnaires to assess latent concepts such as loneliness or extraversion finding the best behavioral or cognitive indicators is often a tedious process. Operationalizations suffer from excessive degrees of freedom in the number of indicators used to measure a construct and their merely probabilistic relationship with the latent variable (Meehl, 1978). In a culture of hypothesis testing, imprecise measurement tools are problematic since it becomes almost impossible to discern relationships between variables based on meaningful variance as opposed to noise (Meehl, 1978). Thus, researchers have advocated to shift the focus to nonconfirmatory activities such as construct formulation and measurement development to make hypothesis testing productive (Bringmann et al., 2022; Scheel et al., 2021).

Two Basic Approaches to Construct Definition

In general, psychology distinguishes between two basic approaches to construct definitions: a quantitative approach, or what we call top-down theory building, and a qualitative approach, or what we call bottom-up theory building. Top-down theory building relies on confirmatory research itself. Drawing from previous empirical evidence and domain knowledge, experts will inform the development of questionnaires, which will be validated through factor analyses and correlational studies (Fried, 2020b). Top-down theory building assumes that experts' knowledge about a construct is exhaustive. Given human biases (Nickerson, 1998) and motivation (Rosenthal, 1994), it is doubtful, however, that even experts can grasp the entire space of possible components of a construct. For example, the most commonly used scales to assess major depressive disorder show a wide range of disparate symptoms with only moderate overlap (Fried, 2017, 2020a). What's more is that current forms of validation such as the multitrait-multimethod approach (Campbell & Fiske, 1994) are weak given that all psychological measures suffer from similar drawbacks (Meehl, 1978). Self-report questionnaires can have favorable factor structures and correlational patterns even if they were deliberately constructed to measure a nonexistent concept (Maul, 2017). On the contrary, empirical evidence indicates that classically constructed questionnaires are extremely noisy since response patterns can be predicted almost entirely from the semantic characteristics of the items alone (Arnulf et al., 2014). This suggests that questionnaire scores depend only little on interindividual variations in the latent variable, but heavily on how items were phrased in the first place.

Another approach to the conceptualization of psychological constructs is qualitative research, or what we call bottom-up theory building. Qualitative research often relies on the manual inspection and coding of text data. Fine and Elsbach (2000) point out that sociologists frequently rely on qualitative studies, while modern psychology often dismisses those methods despite their multiple

advantages. First, qualitative data captures real-world processes richer and more dynamically than quantitative data does. Second, in qualitative research, participants express themselves through natural language, a response mode that is less constrained and conforming with people's habits. Third, qualitative methods are more flexible in the sense that they allow the observation of behavior that is hard to study in the laboratory. However, qualitative methods are not free of doubts in psychology (Bhati et al., 2014; Kidd, 2002) especially since bottom-up research misses qualities important to quantitative scientists like transparency (Moravcsik, 2014). Many researchers consider qualitative methods less attractive due to efforts in data handling and structuring the research process (Pokorny et al., 2018). Despite standardization efforts (Tong et al., 2007), replication of qualitative research is often cumbersome and unreliable. Manual inspection of text is tedious and labor-intensive, strongly limiting the scalability of qualitative methods to large-scale data. This hinders our ability to generalize insights from qualitative research that is constrained to small sample sizes.

Since latent concepts lie at the heart of psychological science, finding new and better ways to achieve valid construct definitions is of utmost importance. In this article, we propose a new method for the formalization of constructs, the construct mining pipeline, which unites advantages and mitigates disadvantages of top-down and bottom-up theory building, leveraging computational methods for text data. More specifically, the pipeline is intended for constructs that are not directly connected to objective physiological measurements (e.g., eye movement in attention), but that manifest in thoughts and behaviors (e.g., most constructs in differential and social psychology, such as personality). Examples include emotion regulation (i.e., influencing emotions) (Gross, 1998), interoception (i.e., perceiving bodily signals) (Vaitl, 1996), counter speech (i.e., opposing hate speech) (Benesch et al., 2016), and love languages (i.e., expressing affection) (Chapman, 1990). All of those constructs share the possibility to build theory either qualitatively or quantitatively.

First, we describe how natural language processing (NLP) has been employed in qualitative research and psychometrics. We argue that NLP cannot only make existing processes more efficient but can potentially result in entirely new methodologies. Next, we introduce the construct mining pipeline and its underlying techniques in a step-by-step guide. Finally, we illustrate the application of the pipeline on an example from emotion regulation (ER), that is how individuals experience, express, and influence their emotions (Gross, 1998). Although taxonomies of ER strategies exist, an integrative classification is missing, preventing research from gaining a comprehensive understanding of ER overall. With the construct mining pipeline, we infer a data-driven taxonomy of ER strategies, which is replicable and scalable. We provide new perspectives on already existing theory and demonstrate how our method potentially unites competing points of view within the same construct.

NLP in Psychology

During the last years, methods from computer science have spilled over to psychology enabling new forms of research designs and knowledge creation. For example, mathematical modeling and network science allow more rigorous theory building (Borsboom et al., 2021), agent-based modeling can explain collective behavior (Smith & Conrey, 2007), and the automatic processing of large-scale observational data has made it possible to gain more ecologically valid support for theoretical claims (Garcia & Rimé, 2019; Pellert et al., 2021). Especially methods from NLP, that is, the automated processing of text data, have been welcomed by the psychological community. A growing body of literature provides guidance for the adoption of NLP in psychological studies (Berger & Packard, 2022). Yet, the idea to apply quantitative methods to text analysis is not new. One of the most commonly used methods in the social sciences are dictionary approaches, such as the Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022). LIWC detects specific concepts (e.g., negative emotions) in text by analyzing the frequency of words related to that concept. Boyd and Pennebaker (2017) and Boyd et al. (2020) have argued that language can be a stable and specific indicator for multiple personality traits, and is more closely related to behavioral outcomes than traditional self-report measures.

Recently, even more sophisticated NLP techniques find application in studies working with text data. A very prevalent application concerns the training of machine learning algorithms to automate human content coding (Wang et al., 2022). It was shown that classification algorithms are suitable for the replication of human-developed coding schemes, although needing a substantial amount of training data. On the other hand, unsupervised techniques such as topic modeling can support the creation of coding schemes, while human oversight will never be obsolete due to the necessity of algorithmic tuning (Nelson et al., 2021). Studies showed that topic modeling and content analysis yield similar results, where topics represent constituting components of themes with higher resolution (Baumer et al., 2017). Other researchers have used quantitative methods to obtain an overview over large amounts of data before proceeding with in-depth qualitative analysis (Andreotta et al., 2019).

Rather than enhancing typical qualitative procedures, other lines of research have attempted to create entirely new indicators for textual data. For example, Pokorny et al. (2018) have introduced network theory for the quantitative analysis of text. Content codes are represented as nodes in a network, and connected through undirected edges when co-occurring in the same piece of text, or through directed edges when following each other. Degree, centrality, or community metrics then allow for a quantitative interpretation of text. This type of analysis further has the advantage to provide neat visualizations for otherwise hardly conceivable data.

Psychometrics, a subdiscipline closely related to construct definitions, has also benefited from NLP. Specifically, machine learning can facilitate the time consuming process of item generation in psychological scale development (Götz et al., 2023; Hommel et al., 2022). Generative language models are deep neural networks that produce fluent, human-like text based on a prompt, that is a short text or sentence (Radford et al., 2019). Götz et al. (2023) show that, given a small number of predefined items, generative language models can produce hundreds of face-valid items in the desired format for new or existing constructs. While items created by the

machine still require expert judgment and selection, the approach makes the initial phase of item generation much more efficient. Since generative models are trained on large corpora of text to recognize meaning patterns, they are likely to propose unexpected, but valuable facets for the measurement of constructs.

Finally, endeavors have been made to find new ways of psychometric assessment based on natural language altogether. Kjell et al. (2019, 2022) have proposed to base psychometric measures on open-ended questions as an alternative to traditional rating scales. Therefore, participants are asked to evaluate their life satisfaction with 10 descriptive words or a short paragraph of text. Using NLP algorithms, numerical representations of those answers are calculated in a way that preserves their semantic meaning. Satisfaction with life scores are obtained by subtracting the “distance” of participants’ answers to words representing dissatisfaction from the “distance” of participants’ answers to words representing satisfaction. Simply put, text answers that are located closer to dissatisfaction result in lower satisfaction with life scores, while text answers that are located closer to satisfaction result in higher satisfaction with life scores. Since semantic measures choose natural text as response mode, they show higher ecological validity compared to traditional rating scales (Kjell et al., 2019). Most importantly, participants are allowed to interpret the construct directly without detouring over researcher’s operationalizations.

Current applications in psychological research demonstrate impressively how NLP can transform the use of text data, especially in psychometrics. State-of-the-art language models help researchers uncover structures in text, providing new perspectives on old questions. Next, we introduce the construct mining pipeline, a new approach to reveal the structure of psychological constructs building on the advantages of combining text data with NLP techniques. In a step-by-step guide, we describe the employed methods and their purpose, the required inputs and resulting outputs for each stage. After a general overview, we provide a detailed example on the application of the construct mining pipeline in the field of ER that provides evidence for the usefulness of the approach and gives further details about the computational methods.

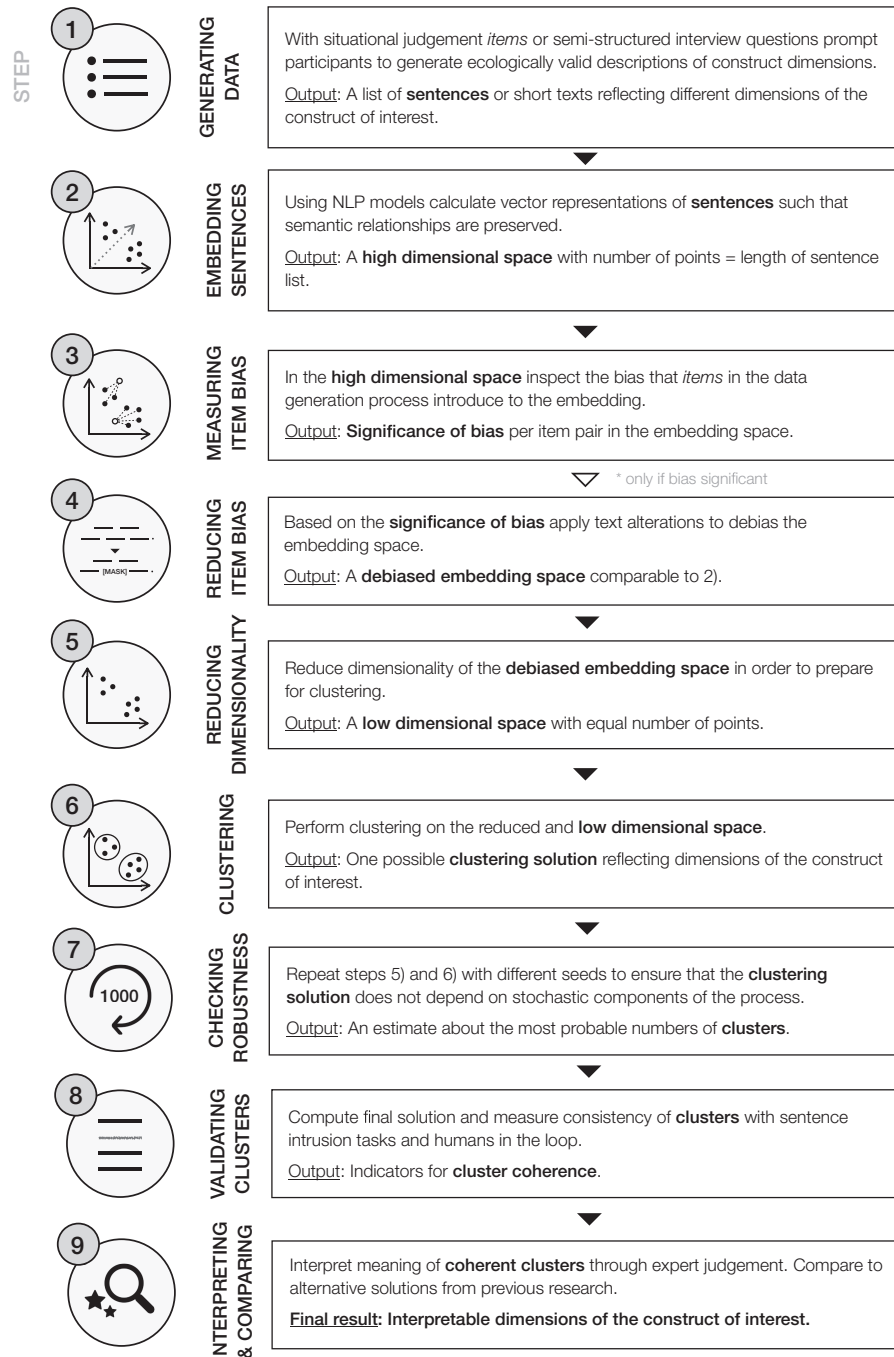
Introducing the Construct Mining Pipeline

We propose a new method to reveal the dimensions of psychological constructs from semistructured text data termed the construct mining pipeline. Compared to what has been referred to as Computational Grounded Theory (Nelson, 2020), the current framework goes beyond the idea of enhancing qualitative analyses with computational tools. We note that the proposed method is not an alternative form of psychological assessment. Rather, the construct mining pipeline supports researchers in the operationalization of psychological constructs through a systematic intertwining of psychological and computational techniques. In other words, we aim to enable researchers to find subcategories of latent variables in a data-driven, yet standardized way. The proposed method unites advantages of bottom-up and top-down theory building, preserving the richness of text while meeting the demands for structured and objective data processing. We provide a comprehensive form of operationalizing psychological constructs with greater independence from researcher’s choices that allows for more precise measurement tools, which can be used to characterize individuals and group differences regarding the construct of interest.

The construct mining pipeline involves nine steps illustrated in Figure 1 and explained in more detail in the following. Importantly, the first step of the pipeline requires (a) data generation. In this step, we tailor the data to prepare for the following computations. We give open-ended questions to participants that are designed to generate sentences each reflecting an instance of the construct of interest. Next, we apply (b) sentence embeddings, a state-of-the-art NLP technique

transforming sentences to points in a numerical space such that their semantic relationship is preserved. This transformation renders computations on the text possible. Along the pipeline, we implement several quality measures to ensure that the discovered construct dimensions are valid, and as independent from researchers' choices as possible. We then measure (c) and reduce item bias (d) if present, to clear the data from unwanted traces of the data generation, for example, from the

Figure 1
The Nine Steps of the Construct Mining Pipeline



Note. NLP = natural language processing

questions used to obtain the sentences describing instances of the construct. We perform (e) dimensionality reduction, that is, a modification of the numerical space to prepare for clustering, and (f) the clustering itself used to infer the dimensions of the construct of interest. We perform a (g) robustness check to ensure that the cluster solution is independent from parameters that require random initialization. For the (h) validation of clusters, we administer a specialized task to people without domain knowledge confirming the semantic meaningfulness of clusters. Phrased differently, we test whether the clustering found by the algorithms makes sense to humans as well. Finally, we (i) interpret and name the clusters representing the dimensions of the construct of interest. This includes discarding or merging classes based on the information from previous steps.

Step 1: Data Generation

The first step is to create one or more open-ended questions that yield sentences each describing an instance of the construct of interest. For example, in the case of ER, sentences are supposed to reflect instances of different ER strategies, that is behaviors or thoughts with which people voluntarily change their current emotional state. Notably, the pipeline is less suitable for constructs that are not well expressed in language, most often basic psychological processes like memory or perception. We discuss the range of application in more detail at the end of the article. Open-ended questions can come in different forms, such as structured interviews, situational judgment tests, or diaries, as long as the questions yield the desired data format, that is, a list of sentences. Researchers should weigh between asking about the construct in a more general fashion (e.g., what does the construct mean for respondents), and asking about the construct in a more specific fashion (e.g., using situational judgment questions). While general questions can result in less linguistic bias (i.e., the wording of the open-ended questions reflecting in participant's answers; see also Steps 3 and 4), more specific questions can make thoughts and behaviors related to the construct more accessible for respondents. Researchers may rely on their domain knowledge and the principles of questionnaire development to create suitable questions. Since the data collection comprises short, open-ended questions, we do not need to rely on classic survey tools only. Instead, applying a variety of tools is possible, for example, using online survey platforms (e.g., Prolific), but also embedding the questions into online environments such as games or apps more naturally. Depending on the tool, samples might be selective. While researchers can explore the construct with a biased sample and expand their data later, representative samples are essential for constructs influenced by sociodemographic variables. Taking economic considerations into account, we recommend administering more than one open-ended question to broadly cover the construct of interest and even out bias that specific questions might introduce to the data. After obtaining the raw sentences, we will transform them into numerical data in the next step of the pipeline.

Step 2: Sentence Embeddings

Now, we obtain numerical representations of the texts generated in Step 1. In contrast to the raw data, numerical representations allow for a quantitative processing of the text. Sentence embeddings provide those representations by taking a list of sentences as input and generating a list of vectors (i.e., an ordered list of numbers of predefined length) as output. In other words, sentence embeddings transform a list of sentences

into an equal amount of points in a high-dimensional space. Sentences of higher semantic similarity such as "I drink water" and "I drink juice" will be located closer in space than "I drink water" and "I pet my cat."

Earlier forms of sentence embeddings were based on co-occurrence matrices of words within and across sentences, such as term frequency-inverse document frequency (TF-IDF) (Sammur & Webb, 2010). However, language representations based on co-occurrences suffer from low generalizability and large memory requirements, and do not consider the context of words, an essential human ability in the understanding of text. Today, the state-of-the-art methods for sentence embeddings are transformer models (Vaswani et al., 2017), which are deep neural networks trained to represent language as numbers while preserving semantic relationships within and between texts. One of the most widely used models is BERT (bidirectional encoder representations from transformers) (Devlin et al., 2019), which exists in over 100 languages and can be adjusted to different applications (e.g., text classification). Transformers are pretrained on large corpora of text such as Wikipedia or Google Books. Pretrained models are freely available on the Hugging Face platform (<https://huggingface.co/>) and suitable to run on modern consumer-grade personal computers. They can be downloaded and used out-of-the-box with a Python package called SentenceTransformers (UKPLab, n.d.). R users can refer to packages *reticulate* (Retrieved March) or *text* (Retrieved March).

For the demonstrated application of the construct mining pipeline, we use a model with BERT as backbone that was specifically trained to represent the similarity of sentences (*gbert-large-sts*; where *g* stands for German and *sts* for semantic textual similarity; similar models exist for other languages such as English). With this model, we transform our list of sentences into points in a high-dimensional space. Based on this numerical representation, we can proceed with further steps such as measuring item bias (Step 3).

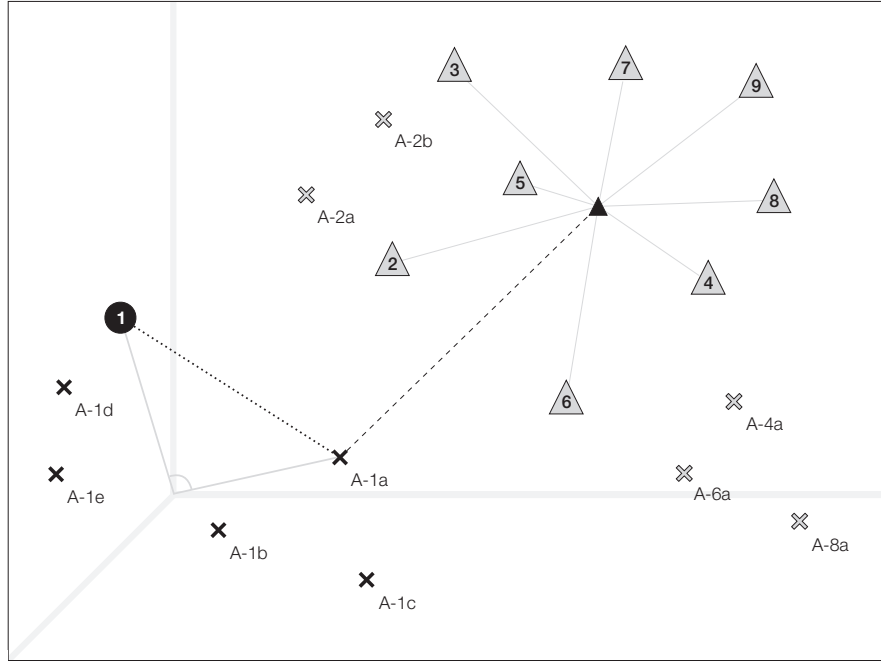
Step 3: Item Bias Measurement

Item bias describes unwanted traces that the data generation introduces to the sentences and hence their embeddings. Embeddings will be influenced by the specific choice of words in a text. The concern is that the phrasing of the open-ended questions (Step 1) might reflect in respondents' answers and, as a result, clusters on the sentence embeddings might emerge for specific terms rather than superordinate construct dimensions. For example, in the case of emotion regulation, clusters might stem from specifics of emotional situations (e.g., flight anxiety) rather than ER strategy classes (e.g., social support).

We propose a procedure to quantify item bias by adapting a method from Caliskan et al. (2017). The authors quantified gender bias in word embeddings by examining whether words for occupations and attributes commonly associated with women are located more closely to words representing the female gender (e.g., "woman") than to words representing the male gender (e.g., "man"), and vice versa. Likewise, we measure whether sentences produced in response to a certain item (i.e., open-ended question) are located more closely to that item than to the other items. We provide a graphic illustration of the procedure in Figure 2.

First, we obtain numerical representations of the items by running them through the same transformer model we used to obtain the sentence embeddings. We calculate the item bias one versus many, that is Item 1 (i.e., open-ended question 1) against Items 2–9, Item 2 against

Figure 2
Illustration of Item Bias Measurement



Note. The item bias statistic reflects whether sentences produced in response to item j (here $j = 1$) are closer to item j in the embedding space than to all other items I (here $I = \{2, 3, 4, 5, 6, 7, 8, 9\}$). We illustrate an existing bias for Item 1 in a 3D space for simplification. All distances are supposed to be interpreted as cosine distances. Sentences produced with respect to Item 1 (A-1a, A-1b, A-1c, A-1d, and A-1e; A for answer) lie closer in space to Item 1 (denoted by the black circle), while sentences produced with respect to Items 2–9 (A-2a, A-2b, A-4a, A-6a, and A-8a) lie closer in space to Items 2–9 on average (individual items denoted by gray triangles, mean denoted by black triangle). Please note that we draw a subset of answers produced with respect to Items 2–9 of equal size to the set of answers for Item 1. 3D = three-dimensional.

Item 1 and 3–9, and so on. Let us consider bias between Item 1 and all other items in the following. The measurement is based on cosine distance as a distance metric. Cosine distance (or cosine similarity) has been used as the standard measure in semantic spaces mainly due to theoretical considerations. Text embeddings are trained such that their outputs reflect linguistic dimensions, although not necessarily interpretable. That said, cosine similarity is high, when two texts have the same proportions of linguistic dimensions (reflected by cosine similarity's nominator, the dot product of two vectors) regardless of the length of the two texts (reflected by the denominator, the product of the vector lengths) (Jurafsky & Martin, 2024; Manning et al., 2009). Since text representations have evolved from bag-of-words approaches to (large) neural networks, researchers are now debating, in which cases cosine similarity might not be favorable as a distance metric (Steck et al., 2024). However, analogies being observed in neural word embeddings still provide strong evidence for embedding dimensions being linguistically meaningful (Bolukbasi et al., 2016). When being tested with different distance metrics in a semantic similarity task, sentence embeddings showed similar performance over all metrics (cosine, Euclidean, and Manhattan) (Reimers & Gurevych, 2019). In any case, cosine distance has the advantage over Euclidean distance of being normalized between 0 and 2 (0 and 1 for text embeddings effectively having only positive dimensions), which makes similarities comparable over different embedding spaces.

For all sentences produced in response to Item 1, we calculate the distance between the sentence and Item 1, as well as the average distance between the sentence and all other items, using cosine distance. Both distances are subtracted and summed up over all sentences produced in response to Item 1. Next, we do the same for a randomly drawn subset of equal size for sentences produced in response to Items 2–9. Finally, we subtract both sum scores to obtain the item bias indicator. An indicator of zero stands for the absence of bias, while indicators deviating from zero suggest the presence of bias. Negative indicators represent bias in the expected direction (i.e., answers of Item 1 are closer to Item 1 than to all other items), while positive indicators represent bias in the unexpected direction (i.e., answers of Item 1 are closer to all other items than to Item 1 itself). Equations 1 and 2 show the exact formula, with J and K being sets of sentences produced in response to item j and all other items k_i , \vec{s} being sentence vectors, \vec{j} being the item vector of item j , and \vec{k}_i being item vectors of all items except j in set I . In order to get a reliable estimate of the magnitude of bias, we repeat the calculation with different sets for J (sentences produced with respect to item j) and K (sentences produced with respect to all other items), where we sample with replacement from J and K , respectively, to obtain two sets of size J . With this empirical distribution, we can calculate the mean and a 95% confidence interval for our bias statistic. If the confidence interval does not contain zero, we

will attempt at reducing item bias in Step 4.

$$\text{ib}(J, K) = \sum_{\vec{s} \in J} d(\vec{s}, \vec{j}, I) - \sum_{\vec{s} \in K} d(\vec{s}, \vec{j}, I), \quad (1)$$

$$\text{where } d(\vec{s}, \vec{j}, I) = \cos(\vec{s}, \vec{j}) - \text{mean}_{\vec{k}_i \in I} \cos(\vec{s}, \vec{k}_i). \quad (2)$$

We can further determine an effect size for our statistic by calculating the difference between the means instead of the sums for sets J and K in Equation 1, and dividing it by the standard deviation of the differences in cosine distances for the combined set of J and K . The formula for the effect size is given in Equation 3. The effect size will help us later in judging whether attempts of reducing bias (Step 4) have been successful. Smaller effect sizes after bias reduction indicate the weakening of the bias effect. Similar to the item bias statistic, we can obtain the mean and 95% confidence interval for the effect size by repeating the calculation with different sets for J and K .

$$\text{es}_{\text{ib}}(J, K) = \frac{\text{mean}_{\vec{s} \in J} d(\vec{s}, \vec{j}, I) - \text{mean}_{\vec{s} \in K} d(\vec{s}, \vec{j}, I)}{\text{SD}_{\vec{s} \in J \cup K} d(\vec{s}, \vec{j}, I)}. \quad (3)$$

The calculation of item bias is only relevant for data that was generated with more than one open-ended question. However, we recommend using more than one open-ended question, because it evens out bias that the way of asking introduces to the data. If item bias statistics deviate from zero, we try to reduce the item bias (Step 4). If statistics are close to zero, we can proceed directly with dimensionality reduction and clustering (Steps 5 and 6).

Step 4: Item Bias Reduction

We try to reduce unwanted traces in the data, if item bias statistics deviate from zero. The simplest, yet most effective, procedure is to mask topic words from the items in the sentences that are unrelated to the construct of interest. Masking means to replace words with placeholders such as “[MASK].” We identify topic words based on our knowledge about the open-ended questions. For example, using situational judgment questions, items will contain situational cues as to where the situation takes place, with whom, et cetera. Topic words most often come in the form of nouns. We can identify topic words by asking “Which topics unrelated to the construct of interest does the question evoke?” Researchers opting for more general questions might not be confronted with the need to reduce bias. However, more concrete questions, such as situational judgment questions, can help respondents to access their knowledge about the construct of interest more effortlessly.

Possible inflections and common variations of the keywords have to be taken into account. Lexicon extension methods (Di Natale & Garcia, 2023) can help finding expressions related to the keywords. A short introduction to lexicon extension methods can be found in Appendix A. Based on the masked data set, we recompute the sentence embeddings from Step 2. Applying the statistic from Step 3, we assess whether masking results in a reduction, or even elimination of bias. Depending on the application, researchers have to decide whether a full elimination or a mere reduction of bias is desirable. For example, in the case of emotion regulation, we do not aim at a full elimination of bias, since there are situation dependent ER strategies (e.g., situational control) only making sense in the context of the item. We will elaborate on the reduction and elimination of

bias in more detail in the discussion. With the embeddings derived from the masked text, we can cluster on the semantic space to reveal the dimensions of the construct of interest (Steps 5 and 6).

Step 5: Dimension Reduction

To prepare our data for clustering, we need to reduce the dimensionality of our numerical representations. Sentence embeddings usually span a large number of dimensions. For instance, embeddings computed with our model of choice (gbert-sts-large) have 1,024 dimensions. The problem is that in high-dimensional spaces, data becomes sparse and, therefore, loses meaning. This is referred to as the curse of dimensionality. When a comparatively small number of points (here points representing sentences) is spread out over many dimensions, all points are approximately equally far away from each other. Hence, clustering yields meaningless solutions collapsing all points into the same cluster. We include an illustration of the curse of dimensionality for the current data set in Appendix B.

We use uniform manifold approximation and projection (UMAP) (McInnes et al., 2020) as a dimension reduction technique. Contrary to other dimension reduction techniques, such as principal component analysis (PCA) (Jolliffe, 2011), UMAP does not optimize for explained variance, but models the data topologically. We deliberately opt for UMAP, since the topological representation mimics the assumption of sentence embeddings that sentences closer in meaning are also closer in space. There are further dimension reduction techniques modeling the data topologically (e.g., Isomap) (Tenenbaum et al., 2000), however, UMAP has a range of other desirable characteristics, too. For further considerations on alternative dimensionality reduction techniques, we refer to the discussion. UMAP is available as Python (McInnes, n.d.) and R package (Konopka, n.d.). It constructs a neighbor graph in the original space, and projects it to lower dimensions. The algorithm defines a hypersphere (i.e., a high-dimensional ball) around each point with a radius based on the density of points in that area. Points with overlapping radii in the high-dimensional space are also more likely to be connected in the low-dimensional space. The density of a point is approximated by the distance to its k -th nearest neighbors. While low values for k urge UMAP to capture the local structure of the data resulting in separated components scattered across the space, high values for k urge UMAP to capture the global structure linking different components within the data (*umap: Uniform ma*). The parameter k will be determined via hyperparameter tuning together with the number of desired dimensions after dimensionality reduction. Hyperparameters are variables that control the model fitting and that are determined by systematically testing different sets of values with respect to selected criteria. We will give more details on hyperparameter tuning in the next section and will demonstrate the process in detail in the example to follow. Further parameters will be set based on theoretical considerations: the minimum distance that UMAP is allowed to place points together, and the metric that is used for distance calculation. Although UMAP possesses a stochastic component, it is possible to choose a random seed at initialization to make results reproducible. UMAP is often used in combination with hierarchical density-based spatial clustering of applications with noise (HDBSCAN), the algorithm we chose for clustering (Step 6).

Step 6: Clustering

Clustering refers to an unsupervised machine learning technique inferring classes from unlabeled data based on its structure. We chose hierarchical density-based spatial clustering of applications with noise as a clustering algorithm (HDBSCAN) (Campello et al., 2013) to reveal the dimensions (i.e., classes) of the construct of interest. HDBSCAN is particularly suited for our purpose, since its assumptions match the characteristics of the problem. Compared to other popular clustering techniques such as k -means (Jin & Han, 2010; Steinley, 2006), agglomerative clustering (Lance & Williams, 1967; Zepeda-Mendoza & Resendis-Antonio, 2013), or Gaussian mixture models (McLachlan & Peel, 2000; Reynolds, 2009), HDBSCAN has the following advantages: first and foremost, it is one of few clustering algorithms today that does not force the user to specify the number of desired clusters beforehand. This is true to our research question (How many classes are there?) and strongly reduces researcher's degrees of freedom. Furthermore, it is able to detect clusters of uneven sizes and shapes. Lastly, it allows points to be left unclustered, which prevents infrequent classes in the sample or unusual formulations of the sentences to corrupt the clustering.

HDBSCAN represents the data as a graph based on the "mutual reachability distance" of two points, which is the maximum of their actual distance (where different distance measures are possible), the distance of point one to its k -th neighbor, and the distance of point two to its k -th neighbor. This method bridges gaps within low-density regions of the graph, and limits the connectedness within high-density regions of the graph. The larger the value of k , the more conservative the clustering will be, in the sense that HDBSCAN detects only high-density regions and leaves more points unclustered (McInnes et al., n.d.). To arrive at a clustering, HDBSCAN successively drops edges of the graph starting from the highest mutual reachability distance while constantly lowering the threshold. The final number of clusters is automatically detected by optimizing the persistence of clusters and the number of points within clusters. A second parameter, which can be set with theoretical guidance, is the minimum number of points one wishes to constitute a cluster. HDBSCAN is available as Python (McInnes et al., n.d.) and R package (Hahsler et al., n.d.).

Similar to UMAP, the value of k will be determined via hyperparameter tuning. We evaluate the choice of hyperparameters for HDBSCAN and UMAP in conjunction by setting parameters with theoretical guidance where applicable, and testing different sets of parameters to find an optimal solution with regard to our evaluation criteria, where theoretical guidance is missing. UMAP's k and number of dimensions after reduction, as well as HDBSCAN's k will be subject to hyperparameter tuning. UMAP's minimum distance and distance metric, and HDBSCAN's minimum cluster size and distance metric are set based on theoretical considerations. Following the argument that the goodness of a clustering has to be judged in the light of the application (Baden et al., 2022), we focus on the number of clusters and the number of unclassified points as evaluation criteria. On the one hand, we can theorize about a reasonable range of number of existing clusters based on how we think about the construct (e.g., there might not be one hundred, but also not merely two ER strategies) and previous evidence from empirical research. If we do not have a rough intuition about the number of clusters, we can always opt for a more conservative clustering solution and discard or merge classes during cluster

evaluation and interpretation (Steps 8 and 9). On the other hand, we want to reduce the amount of unclassified samples as much as possible (Rosenberg & Krist, 2021), since we can assume that each sentence reflects an instance of the construct of interest based on the data generation, and to find an optimal trade-off between a complete and a conservative clustering. Depending on the application, researchers might choose different evaluation criteria. For further details on hyperparameter tuning, we refer to the following example on ER.

Step 7: Robustness Check

UMAP uses random initialization to optimize computation time. Although UMAP is relatively stable across different runs, the results of the dimensionality reduction may vary, and so does to clustering. UMAP allows to fix the random seed used at initialization for reproducibility. We test a number of different random seeds during hyperparameter tuning (i.e., the systematic process of testing values for variables that control the model fitting). However, to ensure that the final clustering is not dependent on the stochastic component of UMAP, we evaluate whether the number of discovered clusters depends on the random seed at initialization.

We generate a set of 1,000 pseudo-random numbers to rerun UMAP and HDBSCAN keeping all other parameters from Steps 5 and 6 constant. We save the number of discovered clusters for each run and report the distribution afterwards. We expect the same number of clusters for the vast majority of runs with only little variance. If the number of clusters shows great variation for different random seeds, we have to assume that the sentences are not distinctive enough to reflect instances of the dimensions of the construct of interest, and go back to overhaul the data generation process. If the number of discovered clusters is fairly identical, we focus on their content, that is we want to identify whether similar points fall into the same clusters across different solutions. Only if the number of clusters and their respective content is fairly identical over different runs, we can assume that the final solution does not depend on the random initialization of UMAP. We compute the adjusted rand index (Hubert & Arabie, 1985; Steinley et al., 2016) for pairs of clusterings to determine their similarity. If adjusted rand indices for pairs of clusterings are low, we have to go back to the data generation. If adjusted rand indices are sufficiently high, we randomly select a seed from the seeds that generate the majority solution (for number of clusters) in order to obtain the final clustering. The adjusted rand index ranges from 0 to 1, where a value of zero is achieved when comparing two random clusterings. In general, values > 0.90 can be regarded as excellent, > 0.80 as good, > 0.65 as moderate and < 0.65 as poor (Steinley, 2004).

Step 8: Cluster Validation

With the completion of the robustness check (Step 7), we attain the final clustering. The clustering reveals groupings of sentences, where each cluster potentially reflects a dimension of the construct of interest. At this point, we would like to interpret the meaning of the dimensions. However, the interpretation of the clustering through the researcher alone can introduce bias.

Therefore, we propose an additional validation check that investigates the meaningfulness of clusters through a human-in-the-loop

approach. We borrow an idea from topic modeling (Chang et al., 2009) to determine what we call cluster coherence. A sample of people without domain knowledge are asked to perform a short survey termed “intrusion task.” Each item of the survey consists of four options, where three options are sentences from the same cluster, while the fourth option is a sentence from a different cluster. Participants are supposed to judge intuitively which of the four sentences does not belong with the others. If the majority of participants is able to identify the “intruders” for a cluster, cluster coherence is high providing evidence for the meaningfulness of the cluster. The majority can be defined by different means: we can either set a theoretically driven cut-off score or define the majority through the empirical coherence distribution. Imagine each question of the intrusion task being answered by four people. A theoretically driven cut-off score could be a minimum of 75% of identified intruders on average per cluster (i.e., three out of four people—the majority—were able to identify the intruders for a cluster on average). An empirically driven cut-off score could be the fraction of identified intruders that is one standard deviation below the mean of fractions of identified intruders over all clusters. We include multiple items per cluster and report the cluster coherence as the average fraction of found intruders per cluster. Clusters with low coherence will be discarded, where low coherence is defined by the aforementioned cut-off scores.

Step 9: Cluster Interpretation

In the final step of the construct mining pipeline, it is the researcher who interprets the set of coherent clusters. We find titles and/or summary descriptions for the clusters by inspecting their constituting sentences. Word clouds can help illustrating the most important words of a cluster, however, we do not recommend solely relying on summaries. As Carlsen and Ralund (2022) emphasize, it is a common misconception that there is solid interpretation without immersion into the data. In the interpretation step, we merge similar clusters. Furthermore, we discard clusters emerging out of other reasons than the construct of interest (e.g., topic words). Researchers are free to invite other experts to interpret the clustering using the experts’ agreement as a further validation check.

Advantages of the Construct Mining Pipeline

We developed the construct mining pipeline as an alternative form of theory building in order to mitigate weaknesses of traditional approaches based on domain knowledge and factor analysis. We do not limit the comprehensiveness with which a construct can be assessed by forgoing predefined rating scale items. Instead, we outsource the verbal description of the construct, which in top-down theory building is performed by a few researchers only, to a sample of nonexperts with carefully tailored open-ended questions. Compared to traditional qualitative analysis, the method we propose is computer-assisted, limiting personal choices during the research process. Furthermore, our method allows to recompute the results or enhance the findings with new data making construct operationalization replicable and transparent. Finally, our approach is scalable. While traditional qualitative analysis is limited to small sample sizes, we can carry out the construct mining pipeline on large amounts of text with a substantial reduction in data processing costs and manual oversight.

The Case of ER

We will demonstrate the construct mining pipeline on an application from emotion regulation research. Emotion regulation is the “process by which individuals influence which emotions they have, when they have them and how they experience and express these emotions” (Gross, 1998). The study of ER often revolves around the different strategies people use in order to regulate their emotions. Early research investigated the usefulness of common strategies with respect to affective experience, interpersonal functioning, and well-being, finding that the habitual use of reappraisal (i.e., the reevaluation of an emotional situation to alter its emotional impact) has benefits over suppression (i.e., the suppression of the emotional expression or the emotion itself) (Gross & John, 2003; John & Gross, 2004). Following research found that the effectiveness of ER strategies depends on situational affordances. For example, reappraisal only has positive effects on well-being if exerted in uncontrollable compared to controllable situations (Haines et al., 2016). Eventually, the concept of flexible ER was introduced, meaning that the effectiveness of a strategy depends on situational characteristics, as well as personal resources and preferences (Aldao et al., 2015; Doré et al., 2015; Kobylińska & Kusev, 2019).

While beneficial and detrimental conditions are well understood for the most common strategies, researchers are still struggling to describe what constitutes overall healthy ER. One reason is the absence of a comprehensive taxonomy for ER strategies. Less prominent strategies have been neglected by experimental research due to their lack of a clear definition and their infeasibility for manipulation in the laboratory (Kobylińska & Kusev, 2019). Without a comprehensive set of ER strategies, researchers are unable to investigate the complex interplay between strategy use, situational affordances and personality traits, core components of ER flexibility (Aldao et al., 2015). A comprehensive strategy taxonomy would further promote the investigation of emotion polyregulation, that is, the simultaneous use of multiple ER strategies within the same emotion episode (Ford et al., 2019).

Strategy taxonomies that exist suffer from disadvantages of top-down theory building being either too broad to be informative or seemingly including an arbitrary number of strategies. Due to their broad categories, theory-driven taxonomies (Folkman & Lazarus, 1980; Gross, 1998) are hardly eligible for studying the intricate interactions of concrete ER behaviors. On the other hand, questionnaires developed to assess ER behavior make implicit or explicit assumptions about a set of ER strategies, while authors hardly ever provide a justification about which strategies are included (e.g., ERQ, Gross & John, 2003; CERQ, Garnefski & Kraaij, 2007; BERQ, Kraaij & Garnefski, 2019; RESS, De France & Hollenstein, 2017; HFERST, Izadpanah et al., 2019; MSCEIT, Mayer et al., 2002; STEM, MacCann & Roberts, 2008; ERP-R, Nelis et al., 2011). A third type of taxonomies comes from qualitative analyses as a byproduct of quantitative research (Southward et al., 2018). These categorizations are comparatively comprehensive, but suffer from drawbacks such as high levels of researcher’s degrees of freedom. Although previous research gives us an intuition about the possible number and space of ER strategies, a taxonomy, which satisfies a range of different quality criteria (Skinner et al., 2003), has yet to be proposed.

Applying the construct mining pipeline, we will infer a data-driven taxonomy of ER strategies, which is scalable, replicable,

and, most importantly, based on descriptions of ER attempts of real people. In the previous sections, we provided an overview over all components of the construct mining pipeline elaborating on the computational methods involved. In the upcoming sections, we will demonstrate how we derived a taxonomy of emotion regulation strategies following the steps of the pipeline. We will close by discussing strengths and weaknesses of our approach, its position within theory building, and its suitability for creating valid and reliable psychological assessment tools.

Using the Construct Mining Pipeline to Develop a Data-Driven Taxonomy of ER Strategies

We illustrate the construct mining pipeline on an example from ER. More precisely, we develop a data-driven taxonomy of ER strategies. We present intermediate results along the nine steps of the pipeline in the following. All data, analysis code, and materials have been made publicly available on GitHub and can be accessed at <https://github.com/Hai-Lina/construct-mining-pipeline>.

Step 1: Data Generation

We used multiple open-ended questions to create a list of sentences each describing a particular ER strategy (i.e., an instance of the construct of interest). Researchers have adopted situational judgment tests (SJTs) common in personnel selection (Lievens et al., 2008) for the measurement of ER. SJTs use short descriptions of hypothetical situations to inquire about participants typical or best behavior constituting an economic trade-off between ecological validity and standardization. One of the most popular instruments to assess ER performance is the situational test of emotion management (STEM) (MacCann & Roberts, 2008). In the STEM, participants are supposed to judge the effectiveness of four regulation suggestions for 44 emotional vignettes in total.

We selected vignettes from the German adoption of the STEM (Hilger et al., 2012) and adapted them for our purposes. We named our adoption emotion regulation strategy inventiveness task (ERSIT). We chose nine out of the 44 original vignettes to ensure good coverage, as well as reasonableness. For standardization purposes, we indicated the emotion that the situation was supposed to evoke at the end of each vignette. Importantly, we balanced characteristics influencing strategy choice over all vignettes, namely the type of emotion (Rivers et al., 2007), controllability (Haines et al., 2016; Troy et al., 2013), emotional intensity (Sheppes et al., 2011), temporal distance (Doré et al., 2015), and social context (Hofmann, 2014). Three items of the ERSIT elicited anger, fear, and sadness, respectively. Considering interactions between strategy, person, and situation, we reasoned that people are only able to show their full strategy repertoire if emotional situations are sufficiently diverse. An example of one of our vignettes is “You find out that some members of your social sports team have been saying that you are not a very good player. You are angry.” Participants were asked to imagine being in the situation and, within three minutes, write down as many ideas as possible to make themselves feel better in the respective situation. The time limit was determined based on previous literature (Weber et al., 2014) and pretesting of the task. Participants answered in the form of sentences such as “I talk to a friend about the situation.” We provide full instructions with an exemplary item and all vignettes in Appendix C.

The sample consisted of $N_p = 113$ students from the University of Graz and the Graz University of Technology. The only requirement for inclusion was to be a German native speaker. The majority of students were enrolled in psychology ($n = 88$). The remaining $n = 25$ students indicated a broad variety of subjects mainly from the engineering and natural sciences. Psychology students were able to obtain a certificate of participation for their studies. Age ranged from 18 to 50 years ($\mu = 21.28$, $SD = 3.51$). 73 of the participants identified as female, 38 as male, one as nonbinary and one participant did not indicate their gender. Accumulated over all vignettes and participants, the ERSIT yielded $N_s = 6,064$ short descriptions of ER strategies. We embedded those sentences with a transformer model in Step 2.

Step 2: Sentence Embeddings

We used Python 3.9.12 for data analysis. For the sentence embeddings, we used the SentenceTransformers package (Version 2.2.0) (Reimers & Gurevych, 2019; UKPLab, n.d.). We chose a German BERT model specifically trained to represent the similarity of sentences (gbert-large-sts) (Risch et al., n.d.) and accessed it via the Hugging Face platform. We ran the ER strategy sentences collected in Step 1 through the model obtaining embeddings of 1,024 dimensions and saved the embeddings to a file for further processing.

Step 3: Item Bias Measurement

Item bias describes unwanted traces that the open-ended questions leave in the sentences and hence their embeddings. Since we used more than one item for data collection, we quantified the bias that the formulation of the emotional vignettes brought to the strategy descriptions of participants. For example, ER suggestions created with respect to the vignette “You find out that some members of your social sports team have been saying that you are not a very good player.” included the term “sports team” themselves. We used our novel statistic for the measurement of item bias (for details please refer to the Introductory part of the pipeline). In Table 1, we present item bias statistics for our data set. All items had certain level of bias to correct, as their 95% confidence intervals did not contain zero. If item bias exists, clusters might form based on specific keywords that the open-ended questions introduced to the answers of participants as opposed to clusters forming based on dimensions of the construct of interest. Thus, we attempted to reduce item bias in Step 4.

Step 4: Item Bias Reduction

Since item bias statistics deviated from zero, we aimed at reducing bias in the data. We inspected each of the emotional vignettes in order to identify topic words. For example, for strategy sentences originating from the vignette describing a situation within a sports team (“You find out that some members of your social sports team have been saying that you are not a very good player.”), we decided to mask the words “sports team,” “team,” and “member.” We list the masked words along with the emotional vignettes in Appendix C.

We used regular expressions (regex) to capture inflections and declensions of the topic words. This accounts for the grammatical gender, cases, and adaptation of verbs in the German language (e.g., “Student,” “Studentin,” and “Studenten”). We replaced all keywords and their variations in the sentences with “[MASK]”

Table 1
Item Bias Statistics and Effect Sizes Before and After Item Bias Reduction With 95% Confidence Intervals

Item	Before		After	
	Bias	Effect size	Bias	Effect size
1	-0.99 [-1.19, -0.79]	-0.35 [-0.42, -0.28]	-0.65 [-0.85, -0.43]	-0.25 [-0.33, -0.16]
2	-1.79 [-1.91, -1.66]	-0.85 [-0.90, -0.80]	-1.53 [-1.66, -1.40]	-0.75 [-0.81, -0.69]
3	-2.92 [-3.19, -2.67]	-0.75 [-0.81, -0.69]	-2.40 [-2.66, -2.13]	-0.66 [-0.72, -0.59]
4	-1.96 [-2.10, -1.81]	-0.87 [-0.92, -0.82]	-1.16 [-1.30, -1.01]	-0.50 [-0.56, -0.45]
5	-2.11 [-2.26, -1.97]	-0.91 [-0.96, -0.86]	-1.99 [-2.14, -1.84]	-0.85 [-0.90, -0.80]
6	-0.91 [-1.01, -0.80]	-0.62 [-0.66, -0.56]	-0.69 [-0.79, -0.58]	-0.50 [-0.58, -0.43]
7	-0.86 [-1.00, -0.71]	-0.43 [-0.51, -0.36]	-0.68 [-0.82, -0.53]	-0.36 [-0.44, -0.27]
8	-1.23 [-1.37, -1.09]	-0.63 [-0.69, -0.56]	-1.07 [-1.21, -0.93]	-0.57 [-0.63, -0.50]
9	-1.51 [-1.71, -1.29]	-0.53 [-0.60, -0.45]	-1.49 [-1.69, -1.27]	-0.53 [-0.60, -0.45]

Note. 95% confidence intervals for item j were calculated by drawing random subsets for sentences of item j and all other items of the size of the subset of sentences of item j .

and recomputed the embeddings with the masked sentences. We saved the embeddings to a file and proceeded with the updated embeddings. Table 1 shows the item bias statistics for the embeddings derived from the masked text. Although we did not eliminate all bias, we can observe that all values moved closer to zero (i.e., an unbiased state). With the exception of items 5 and 9 (minus 5% and 1%), we achieved a reduction of 12%–41% in bias and a reduction of similar magnitude in effect sizes. Notably, bias statistics for items 2, 3, 4, and 6 are outside the confidence intervals of the measurements before correction, and confidence intervals for effect sizes before and after reduction are only marginally overlapping (except for item 9), indicating a substantial reduction in bias. We believe that a full elimination of bias is undesirable, since there are situation-specific ER strategies that only make sense in the context of the item, which we will discuss in more detail when interpreting the clusters (Step 9). For a more vivid illustration of the effects of bias reduction, we note that, had we proceeded with the original embeddings, we would have found 12 out of 38 clusters based on topic words, whereas with the embeddings derived from the masked text we only find six out of 37 clusters based on topic words. Similarly, the original embeddings would have resulted in six non-interpretable clusters, whereas the embeddings of the masked text result in only three noninterpretable clusters.

Steps 5 and 6: Dimension Reduction and Clustering

Steps 5 and 6 are the core of the construct mining pipeline, where we compute the actual clustering. We use UMAP (uniform manifold approximation and projection) (McInnes et al., 2020) as dimension reduction technique and HDBSCAN (Campello et al., 2013) as clustering algorithm (see Introductory of the pipeline for details). Both UMAP (McInnes, n.d.) and HDBSCAN (McInnes et al., n.d.) are available as Python packages, where we used Versions 0.5.3 (UMAP) and 0.8.28 (HDBSCAN). Similar to other dimension reduction and clustering algorithms, UMAP and HDBSCAN require the user to set hyperparameters, that is, parameters that are determined by the user to control the model fitting process. Many of the hyperparameters can be deduced from theoretical considerations, while others have to be defined by trying out different sets of values, a process called tuning. A classic approach to tuning is a grid search, where for each parameter a set of values is defined. For each possible combination of parameters, solutions are

examined with respect to specified evaluation criteria. We will first elaborate on the parameters set by theoretical considerations (which are UMAP’s minimum distance d_{min_UMAP} , and distance metric m_{UMAP} , as well as HDBSCAN’s minimum cluster size $cmin_{HDBSCAN}$, and distance metric $m_{HDBSCAN}$) and turn to hyperparameter tuning afterwards (including UMAP’s k neighbors k_{UMAP} and components n_{UMAP} , as well as HDBSCAN’s k neighbors $k_{HDBSCAN}$). Table 2 lists all necessary hyperparameters for UMAP and HDBSCAN.

For UMAP, we chose cosine similarity as distance metric ($m_{UMAP} = \text{“cosine”}$), a common measure for the closeness of two embeddings (in our case two sentences). For minimum distance, we chose $d_{min_UMAP} = 0.01$. We expected close to identical formulations for some strategy sentences within our sample, thus choosing a small value for the minimum distance between two embedded points. For instance, many people might suggest to “go and talk to a friend” as an ER strategy. For HDBSCAN, we adopted Euclidean distance as distance metric ($m_{HDBSCAN} = \text{“Euclidean”}$). We can assume that the Euclidean distance is meaningful after reducing the dimensionality of our embedding space and we would not be able to argue that the dimensions after reduction necessarily represent linguistic dimensions (which would justify the application of cosine distance). For minimum cluster size, we chose a value of $cmin_{HDBSCAN} = 10$ ruling out that only extremely infrequent ER suggestions will constitute their own class.

After setting the theoretically deduced parameters, we turned to hyperparameter tuning. We will tune UMAP and HDBSCAN in conjunction as we evaluate solutions mainly based on the number of clusters they produce and the points left unclustered. We decided against the use of application-agnostic measures (e.g., density-based cluster validity index) (Moulavi et al., 2014), since we wanted to emphasize evaluation criteria with a solid connection to the research question at hand (Baden et al., 2022). We evaluated UMAP embeddings with $n_{UMAP} = 10, 30, 50$ and a fixed number of $k_{UMAP} = 30$ in combination with values for HDBSCAN $k_{HDBSCAN} = 10, 15, 20, 25, 30, 35, 40, 45, 50$. Likewise, we evaluated UMAP embeddings with $k_{UMAP} = 15, 30, 45$ and a fixed number of $n_{UMAP} = 30$ in combination with values for HDBSCAN $k_{HDBSCAN} = 10, 15, 20, 25, 30, 35, 40, 45, 50$. We repeated those steps for four randomly generated UMAP random seeds $random_{UMAP}$. Figure 3 depicts the results for the evaluation criteria number of clusters and number of unclustered points for four different UMAP random seeds (colored lines).

Table 2
Hyperparameters Used to Control the Training Process of UMAP and HDBSCAN

Hyperparameter	Interpretation	Default value	Final value	Obtained through
UMAP				
k_{UMAP}	Number of k neighbors used for density estimation of a point	15	30	Tuning
$d_{\text{min}}_{\text{UMAP}}$	Minimum distance that two points have to be apart after embedding in the low dimensionality space	0.1	0.01	Theory
n_{UMAP}	Dimensions of the low dimensionality space	2	30	Tuning
m_{UMAP}	Metric used for calculating the distance between two points	“Euclidean”	“cosine”	Theory
$random_{\text{UMAP}}$	Random number for initialization of the model (improves speed and efficiency of computations)	None	86,531	Robustness check (Step 7)
HDBSCAN				
k_{HDBSCAN}	Number of k neighbors used for density estimation of a point	None	30	Tuning
$cm_{\text{min}}_{\text{HDBSCAN}}$	Minimum number of points a cluster has to have	5	10	Theory
m_{HDBSCAN}	Metric used for calculating the distance between two points	“Euclidean”	“Euclidean”	Theory

Note. UMAP = uniform manifold approximation and projection; HDBSCAN = hierarchical density-based spatial clustering of applications with noise.

We can observe an overall decreasing trend for the number of discovered clusters with higher values for UMAP’s number of components n_{UMAP} or number of neighbors k_{UMAP} . Similarly, we can observe increasing noise with increasing number of components n_{UMAP} and number of neighbors k_{UMAP} . Some unstable solutions exist indicated by sudden drops in the curves for number of clusters and noise. Especially, for $k_{\text{UMAP}} = 45$ and $k_{\text{HDBSCAN}} = 20$, we can see that the cluster solution collapses into two large clusters with zero points being left unclustered. We expected the number of clusters to lie somewhat around 30 (Southward et al., 2018), so a value of $n_{\text{UMAP}} \geq 25$, as well as a value of $k_{\text{UMAP}} \geq 25$ appear reasonable. We decided to prefer cluster solutions with greater compared to smaller number of clusters, since we expected to obtain a more differentiated picture with the possibility to clean the clustering in the remaining steps of the construct mining pipeline. Furthermore, with increasing number of components n_{UMAP} and number of neighbors k_{UMAP} , noise increases as well, falling together with solutions with smaller number of clusters. We note that there is a considerable amount of noise for all nontrivial clusterings. We suspect that unclassified samples can be attributed to unusual formulations rather than undetected ER strategy classes, and provide an analysis of unclassified samples as a validation check in Appendix D. Inspecting curves for other random UMAP seeds and adding up all considerations, we aimed for the supposedly most stable solution (i.e., not collapsing into two large clusters) with a number of clusters around 30. This left us with a set of hyperparameters of $k_{\text{UMAP}} = 30$, $n_{\text{UMAP}} = 30$, and $k_{\text{HDBSCAN}} = 30$. The value for the random seed of UMAP will be determined in Step 7: the Robustness Check section.

Step 7: The Robustness Check

The current step is supposed to ensure that the discovered cluster solution does not depend on the random initialization of UMAP. We, therefore, repeated the dimensionality reduction and clustering with 1,000 random seeds for the same set of hyperparameters. The goal is to examine whether solutions with different random seeds result in a similar number of clusters. Figure 4 illustrates the distribution of the number of discovered clusters over 1,000 different UMAP and HDBSCAN runs. The distribution is narrow ($SD = 1.42$) with a

majority solution of $n_C = 37$ clusters (for $n_{\text{sol}} = 234$ number of solutions). There is a considerable amount of solutions ($n_{\text{sol}} = 162$), where points collapse into two large clusters. We will neglect those clusterings moving forward, since trivial solutions are uninformative.

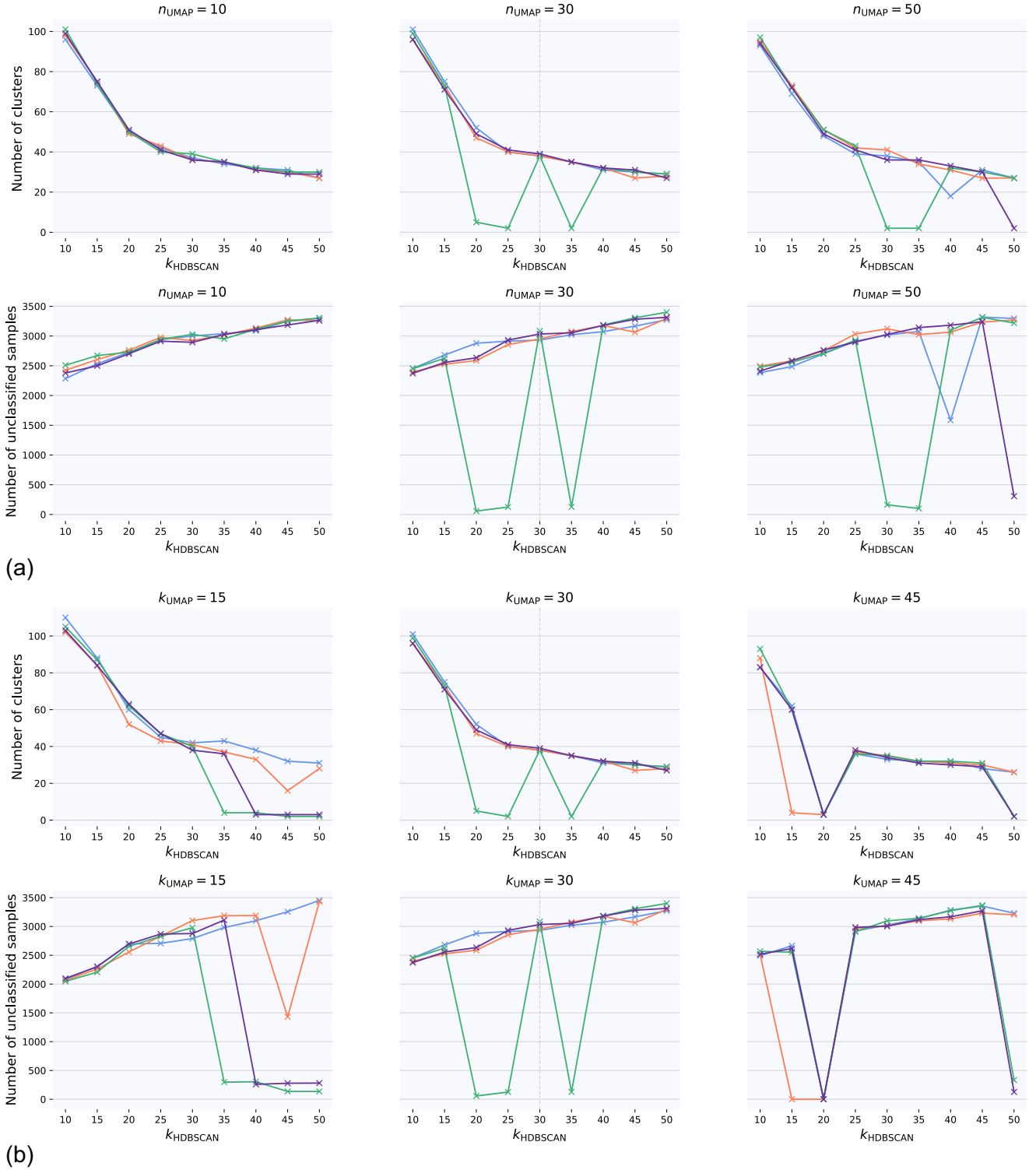
After excluding trivial solutions, we computed the adjusted rand index between pairs of clusterings in order to examine whether solutions overlap in content as well. Adjusted rand indices for pairs of clusterings ranged from 0.65 to 0.88 leading us to conclude that clusterings are overall stable in size and content. We then drew a random seed yielding $n_C = 37$ clusters to determine the final solution. For visualization purposes, we obtained a two-dimensional representation of the embedding space using UMAP ($n_{\text{UMAP}} = 2$), while keeping all other hyperparameters constant. Figure 5 depicts the final solution.

Another possible approach for choosing the final cluster solution is to select the clustering that is most similar to all other clusterings in terms of adjusted rand index. Therefore, we would need to average over adjusted rand indices for pairs of clusterings per random seed. We would then choose the random seed that yields the maximal average adjusted rand index. In our case, the majority solution for number of clusters is fairly similar to the solution with the greatest overlap in terms of adjusted rand index ($ARI = 0.82$ vs. $ARI = 0.84$), and even closer to the overall mean of average adjusted rand indices (after excluding trivial solutions; $\bar{x}_{ARI} = 0.82$). Furthermore, the maximum of the adjusted rand indices might not be representative and depend on the sample of our robustness check. Generally, average adjusted rand indices per random seed follow and even narrower distribution than indices for pairs of solutions as depicted in Figure E1 in Appendix E.

Step 8: Cluster Validation

In order to evaluate whether the clusters identified with computational methods are interpretable by humans, we conducted a survey termed intrusion task (for details please refer to the Introductory part of the pipeline). Including nonexperts in the interpretation of the cluster solution adds another level of objectivity since they will not suffer from confirmation bias.

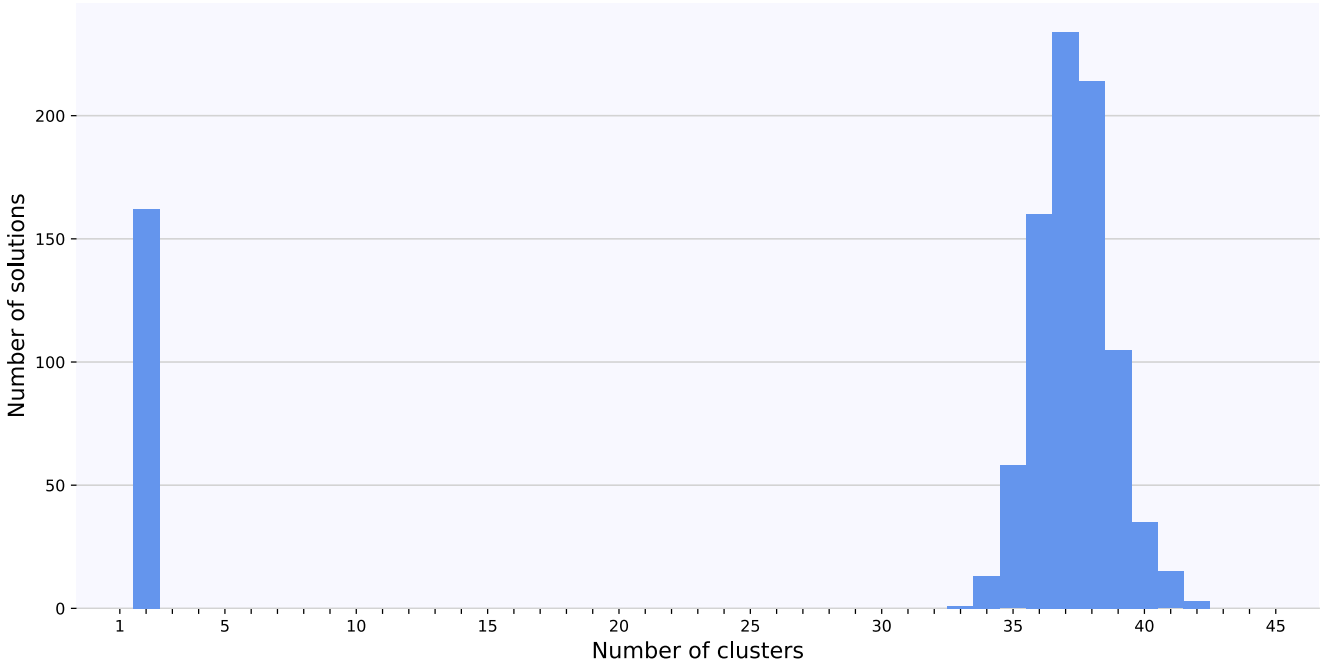
We aimed at evaluating 30% of observations in each cluster meaning that each cluster will contribute an unequal amount of

Figure 3*Evaluation Criteria Over a Variation of UMAP and HDBSCAN Parameters for Four Different Random Seeds*

Note. (a) Number of clusters and unclassified samples over a variation of n_{UMAP} and k_{HDBSCAN} , $k_{\text{UMAP}} = 30$ (b) Number of clusters and unclassified samples over a variation of k_{UMAP} and k_{HDBSCAN} , $n_{\text{UMAP}} = 30$. Different line colors represent different random seeds. The solution for $n_{\text{UMAP}} = 30$, $k_{\text{UMAP}} = 30$, and $k_{\text{HDBSCAN}} = 30$ appears to be the most stable with no dips in the curves, that is, no trivial solutions with only few numbers of clusters. UMAP = uniform manifold approximation and projection; HDBSCAN = hierarchical density-based spatial clustering of applications with noise. See the online article for the color version of this figure.

Figure 4

Distribution of the Number of Clusters Over 1,000 UMAP and HDBSCAN Runs Initialized With Different Random Seeds



Note. A total of 162 random seeds result in trivial solutions, where all points fall into two large clusters, and which can be neglected. Mode = 37, $SD = 1.42$. UMAP = uniform manifold approximation and projection; HDBSCAN = hierarchical density-based spatial clustering of applications with noise. See the online article for the color version of this figure.

items to the survey. For example, Cluster 5 contains $n_s = 174$ sentences resulting in $174 \cdot 0.3 \approx 54$ sentences to be evaluated (we rounded up to next higher number divisible by three). Since one survey item covers three sentences from the same cluster accompanied by an intruder, we divide this number by 3 yielding $54 \div 3 = 18$ items for Cluster 5 in the intrusion task. An exemplary item reads as follows (translated from German): “Which of the following sentences does not belong to the others? A: Calming down B: Meditating/Relaxing C: I train relaxation techniques D: I complain.” This resulted in 316 items in total over all clusters. In order to limit the effort for participants, we restricted the survey to approximately 50 items per person. Accordingly, we had six sets of nonoverlapping item groups ($2 \cdot 52$ items + $4 \cdot 53$ items = 316 items). Every item group was judged by $n_p = 4$ people with a total sample of $N_p = 24$ people. The survey lasted 10–15 min per person. We calculated the average fraction of found intruders over all items of a cluster (Figure 6).

Only three clusters (1, 9, and 23) fall below a threshold of 0.75. That means that less than three out of four people were able to find the intruder for that cluster on average. It might be that for uninterpretable clusters subtle sentence characteristics such as the use of subjunctives dominate with clusters forming based on that. We discard the clusters falling below the threshold for the final step of analysis, the interpretation of clusters.

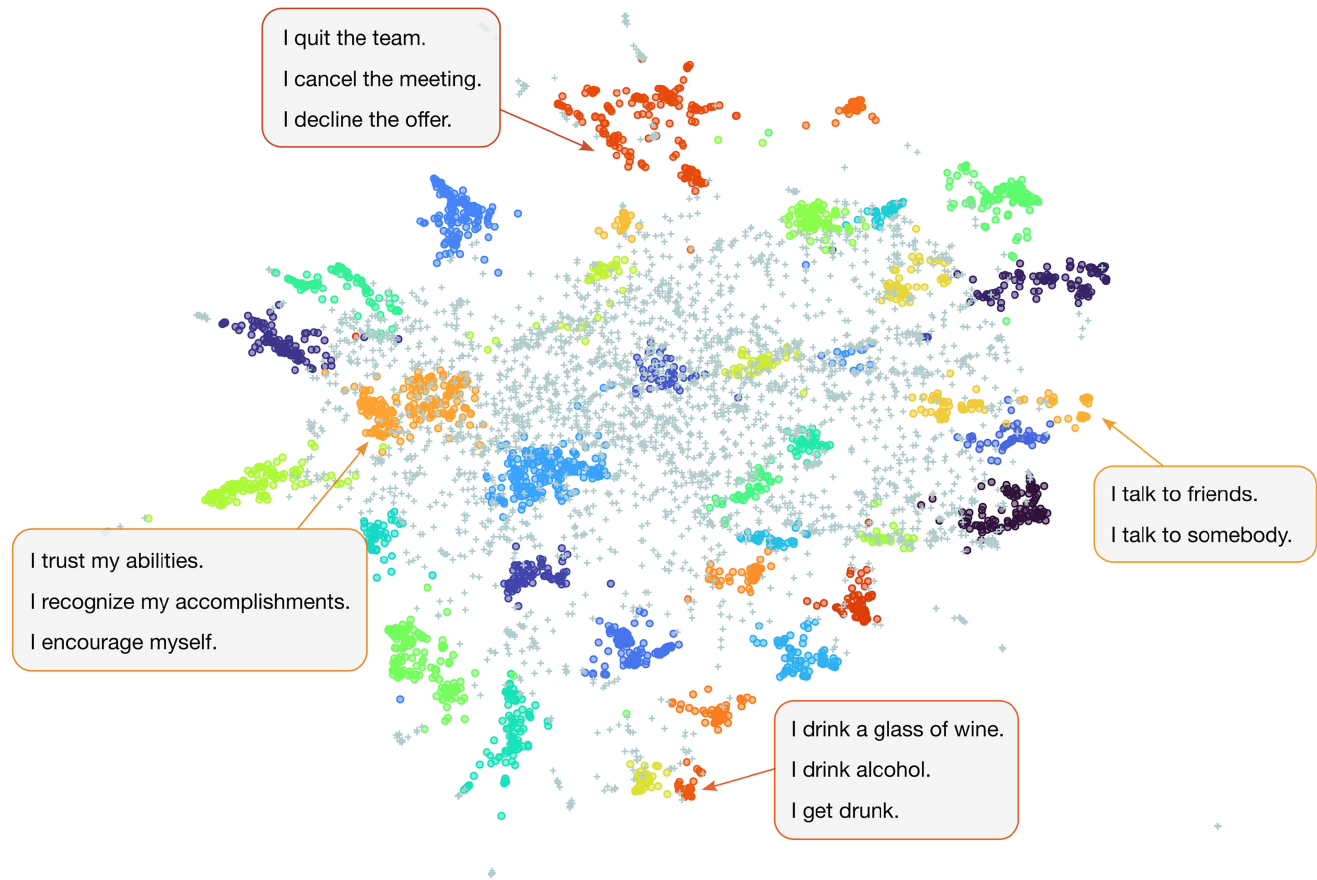
Step 9: Cluster Interpretation

In the final step of our analysis, we interpreted the meaning of clusters by inspecting sentences within them. In order to visualize the

interpretation for the reader, we provide 10 characteristic words for each cluster in Table 3. We computed the topmost frequent words for each cluster by normalizing each word by its frequency within a cluster and its frequency over all clusters (Grootendorst, 2022).

While three clusters were discarded based on low scores in the intrusion task (Step 8), we eliminated another six clusters realizing that those clusters formed due to topic words rather than superordinate ER strategies. Cluster 4 centered around the word “to fly” (ger.: fliegen), Cluster 11 around the word “to call on the phone” (ger.: anrufen), Cluster 14 around the word “to make friends with” (ger.: anfreunden), Cluster 16 around the word “coffee shop” (ger.: Café), Cluster 19 around the word “fear” (ger.: Angst), and Cluster 33 around the word “to complain” (ger.: beschweren). Cluster 35 was not interpretable even though it had a high intrusion score of 0.80. Although the possibility exists that the high score emerged randomly, we hypothesize that people were able to recognize the intruder for Cluster 35 in large parts due to the characteristics of its sentences that have been very different from those of the intruder (i.e., length of sentence). All other clusters were interpretable as ER strategy classes and are summarized in Table 4.

We found 15 strategy classes in total with some classes represented by multiple clusters merged into one. The fact that we find more classes than prompts demonstrates that the taxonomy is somewhat independent from the data generation. Interestingly, the eight clusters contained in “situational control” mainly refer to situational control strategies with respect to different items. For example, when people were asked to imagine taking a plane although suffering from flight anxiety, they proposed to look for alternative means of transport. The two clusters of the class “substance abuse” refer to medication

Figure 5*Final Cluster Solution After Hyperparameter Tuning and Robustness Check*

Note. Exemplary sentences for illustration of selected clusters. Gray points (crosses) represent unclassified samples. Hyperparameters—UMAP: $k_{\text{UMAP}} = 30$, $d_{\text{minUMAP}} = 0.01$, $n_{\text{UMAP}} = 30$ ($n_{\text{UMAP}} = 2$ for illustration), $m_{\text{UMAP}} = \text{"cosine"}$, $\text{random}_{\text{UMAP}} = 85,631$; HDBSCAN: $k_{\text{HDBSCAN}} = 30$, $\text{cmin}_{\text{HDBSCAN}} = 10$, $m_{\text{HDBSCAN}} = \text{"Euclidean"}$ UMAP = uniform manifold approximation and projection; HDBSCAN = hierarchical density-based spatial clustering of applications with noise. See the online article for the color version of this figure.

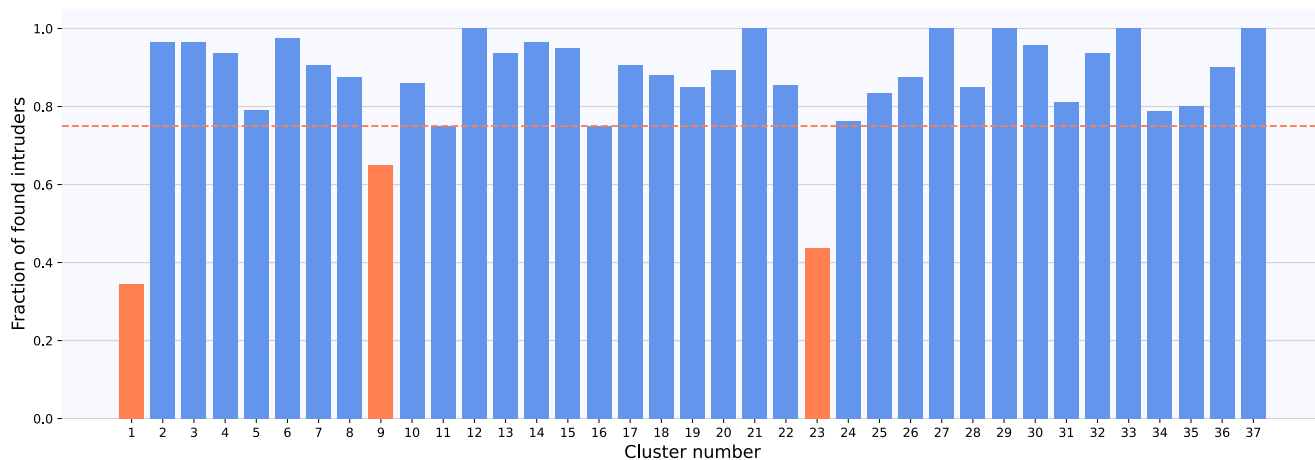
and the abuse of legal drugs such as alcohol. Some clusters formed with respect to a specific item, although representing an overall ER strategy class. For instance, when people imagined that close relatives are sick in the hospital, they proposed to gather further information about their illness, which we termed “information seeking.” This underpins our assumption that a multitude of situations with varying characteristics is needed for people to show their full ER repertoire. For other classes such as “withdrawal,” we see a satisfying mixture of items. For example, people proposed to “quit their sports team,” “change their job,” “look for new friends,” et cetera. We did not only find behavioral, but also cognitive ER strategies such as “self-enhancement” and “reappraisal” represented by formulations such as “I tell myself that ...,” or “I keep thinking that ...” For a full description of all discovered strategy classes refer to Table 4.

General Discussion

We proposed a new method for the data-driven definition of psychological constructs called the construct mining pipeline, which interleaves psychological and computational methods. The method

aims at mitigating disadvantages both from top-down and bottom-up theory building by combining their advantages. The pipeline preserves the richness of qualitative text data, while offering methods for systematic data analysis known from quantitative approaches. We demonstrated the method on an example from ER by developing a data-driven taxonomy of ER strategies. We collected descriptions of ER attempts with nine open-ended situational judgment questions and obtained numerical representations of the sentences using sentence transformers. We measured and reduced traces of the data collection in the embeddings. We clustered on the semantic space to derive superordinate ER strategy classes and examined the robustness of the solution with respect to the stochastic component of the algorithms applied. We assessed the coherence of clusters in a standardized survey using a non-expert sample. Finally, we interpreted the clusters and formed classes of ER strategies. The current data set yielded 15 strategy classes, some of which are well-known to the literature (e.g., suppression), some of which received little to no attention in psychological research (e.g., self-enhancement).

Since the current data set is largely restricted in size and diversity of participants, further research is needed to arrive at a

Figure 6*Fraction of Found Intruders Over All Clusters*

Note. Clusters 1, 9, and 23 fall below the majority threshold of 0.75 (three out of four people) indicated by the dashed line. See the online article for the color version of this figure.

domain-specific conclusion. This work represents a methodological contribution and should be regarded as a proof of concept. Our approach is scalable meaning that further data can be collected and processed to complement earlier findings. Moreover, our approach is replicable, that is results can be recomputed and the same type of analysis can be performed on other data sets. The pipeline is data-driven while strongly limiting researcher's degrees of freedom. Compared to traditional rating scale development, where researchers formulate items based on their understanding of the construct, the pipeline relies on laypersons' descriptions and direct interpretation of the psychological process. Since participants are less limited in their natural expression, this results in a more ecologically valid measure of psychological constructs.

The next step in the development of a data-driven taxonomy of ER strategies is to validate the current findings with a larger, more diverse sample potentially including different languages. Although English is the dominating language in NLP, sentence similarity models exist not only for German, but also for French, Spanish, and Portuguese. With the help of a taxonomy of ER strategies we can infer individual profiles, which can be related to outcomes such as depressiveness, dispositional affect or loneliness. Using the taxonomy to predict different criteria is important to not solely rely on face validity and interpretability of the findings (Baden et al., 2022). A comprehensive taxonomy might promote the investigation of flexible emotion regulation (Aldao et al., 2015) or polyregulation (Ford et al., 2019). Future ER research could further make use of social media data by automatically extracting relevant strategies from text applying the classes of the taxonomy. This approach provides a more ecologically valid perspective on people's thoughts and behavior, while coming at lower cost than traditional approaches such as experience sampling or large panel designs.

The pipeline employs UMAP (McInnes et al., 2020) and HDBSCAN (Campello et al., 2013) as dimensionality reduction and clustering techniques. UMAP is especially suitable for this application because it models the data topologically, which reflects the characteristic of transformer models to represent text such that

sentences closer in meaning are closer in space. Other dimensionality reduction techniques such as PCA (Jolliffe, 2011) are less ideal, because they optimize for different criteria (e.g., variance explained) that are not directly connected to the functioning of sentence embedding models.

In general, dimensionality reduction techniques can be divided along two axes: optimizing for local versus global structure, and applying linear versus nonlinear optimization (Silva & Tenenbaum, 2002; Xia et al., 2021). Another popular dimensionality reduction technique that models data topologically is Isomap (isometric mapping) (Tenenbaum et al., 2000). Similar to UMAP, it maps the data as a graph based on a point's k nearest neighbors, but with two points' mutual Euclidean distance as edge weights for all points within the given radius. The projection is achieved by preserving the shortest path between two nodes (an approximation of their geodesic distance) in the lower dimensional space. While Isomap optimizes the global data structure (Silva & Tenenbaum, 2002), UMAP tends to optimize local structures, but the degree to which it balances local and global structures can be adjusted with its parameter k (McInnes, n.d.). It has been shown that nonlinear and local techniques are preferred for cluster and membership identification, at least in terms of visual cluster evaluation through humans, where UMAP and t-SNE (t-distributed stochastic neighbor embedding) (van der Maaten & Hinton, 2008) perform best out of this category (Xia et al., 2021). Isomap has been adapted to perform advantages generally associated with local methods (C-Isomap or conformal Isomap, and L-Isomap or landmark Isomap) (Silva & Tenenbaum, 2002), but it comes at the cost of additional assumptions, especially data being uniformly sampled from the manifold. In contrast, UMAP enforces the assumption of data being uniformly distributed, although the actual data may be not, which is justified by its mathematical underpinnings (McInnes, n.d.). Additionally, rather than making binary decisions whether to connect two points on the manifold (like Isomap), UMAP assigns a connection probability to pairs of points applying fuzzy topology. Together with a local connectivity constraint ensuring that each point is at least connected to its nearest neighbor, UMAP has the

Table 3

Top 10 Characteristic Words for Each Cluster Normalized by the Frequency of Words Within and Across Clusters

Cluster	Intrusion score	Top words
1	0.34	About, for, about, could, talk, make friends, talk, can, hear, work
2	0.96	Medicine, take, sedative pills, sedatives, take, pills, get, sedative, flight, sedative drops
3	0.96	Train, alternatives, search, means of transport, boat, train, take, transport facilities, location, search
4	0.94	Fly, flight, fly, anxiety, statistics, nevertheless, person, safety, inform, just
5	0.79	Distract, distract, distract, distract, music, distraction, movie, sport, watch, watch
6	0.97	Relax, meditate, break, relaxation techniques, relaxation exercise, deep, breath, do, breath, breathing exercise
7	0.91	Therapy, fear of flying, therapist, go, go, overcome, psychologist, psychological, treat, try
8	0.88	Visit, hospital, visit, relative, drive, relatives, visit, drive, immediately, person
9	0.65	Time, fun, finish, useful, things, do, meanwhile, do, hobbies, things
10	0.86	Ignore, forget, ignore, try, ignore, try, carry on, suppress, about, thoughts
11	0.75	Call, call, contact, telephone, telephone, friend, colleague, stay, best, regularly
12	1.00	Eat, eat, chocolate, ice cream, good, coffee, go, drink, cook, good
13	0.94	Drink, alcohol, glass, get drunk, drink, wine, beforehand, cry, relaxation, beer
14	0.96	Make friends, friends, colleagues, friendships, new, friends, make friends, meet, make friends, try
15	0.95	Meet, date, day, colleague, ask, fix, find, nevertheless, would like to, new
16	0.75	Coffee shop, go, go, Fridays, alone, alone, coffee, colleague, more, invite
17	0.91	Illness, inform, inform, inquire, treatment options, doctor, chances of recovery, investigate, ill, take a close look at
18	0.88	Decline, change, change, job, team, decline, contract, studies, search, search
19	0.85	Fear, fears, speak, family, friends, speak, fears, speak, nervousness, friends
20	0.89	Prepare, plan, day, well, plan, prepare, do, to, list, plan
21	1.00	Grief, cry, cry, let, free, vent, sadness, feelings, hide, feelings
22	0.85	Help, advice, please, help, ask, support, friends, please, colleagues, take
23	0.44	Time, present, beautiful, relatives, person, spend, experience, moments, memories, times
24	0.76	Abilities, strengthen, own, eyes, conscious, good, more, experience, successes, own
25	0.83	Harder, train, more, improve, train, better, exert, try, at, performance
26	0.88	Study, study, more, study groups, tutoring, study group, lectures, form, tutorial, enter
27	1.00	Schedule, progress report, revise, adapt, change, correct, new, new, optimize, restructure
28	0.85	Speak, about, friends, somebody, someone, speak, talk, about, vent, someone
29	1.00	About, speak, talk, friends, friend, friend, tell, about, good, colleague
30	0.96	Family, speak, friends, about, speak, relatives, somebody else, trusted, person, call
31	0.81	Information, acquire, information, myself, get, try, info, just, try, important
32	0.94	Anger, let, rage, frustration, let out, sport, air, at, box, let out
33	1.00	Complain, complain, with, boss, colleagues, friends, responsible, about, superior, get upset
34	0.79	Bad, fault, good, for, tell, realize, tell, perfect, clear, yet
35	0.80	Report, complaint, responsible, reason, person, ask, client, who, point out, client
36	0.90	Boss, situation, superior, explain, matter, about, talk, superior, clarify, report
37	1.00	Boss, trainer, talk, address, about, express, discussion, again, discuss, inform

Note. Clusters 1, 9, and 23 fall below the majority threshold in the intrusion task (Step 8). Clusters 4 (“to fly”), 11 (“to call on the phone”), 14 (“to make friends with”), 16 (“coffee”), 19 (“fear”), and 33 (“to complain”) formed based on topic words. Cluster 35 was not interpretable regardless of a high score in the intrusion task. Words appearing double in a cluster can be traced back to the translation from German. A table with the original German top words is provided in Appendix F.

following advantages over Isomap (and other methods from the family of nonlinear dimensionality reduction techniques): (a) it ensures that no point is isolated in neither the high-dimensional, nor the low-dimensional space, (b) the graph and its projection are not overly connected in dense regions, and sufficiently connected in sparse regions, (c) UMAP focuses on differences in distances among neighbors rather than absolute distances, which represents high dimensional data more adequately, where distances are large, but also more similar. Other weaknesses of Isomap include it being sensitive to noise and outliers (Balasubramanian & Schwartz, 2002), and—compared to UMAP—computationally intensive for large datasets (McInnes, n.d.). Ultimately, single techniques of the construct mining pipeline are not set in stone and researchers are free to experiment with other

methods, if they are deemed to model specific problems more appropriately. Comparing different methods for parts of the pipeline on the classification outcome can also be subject to future empirical studies. For an overview of popular dimensionality reduction techniques see Xia et al. (2021).

HDBSCAN, too, mirrors assumptions about our problem and data. First of all, it does not force the user to predefine the number of expected clusters, an essential part of our research question. Second, HDBSCAN is able to detect clusters of uneven sizes that are not necessarily Gaussian-shaped reflecting the fact that we can expect some classes to be much more frequent than others (e.g., some ER strategies such as situational control might be more prevalent than others such as self-harm). Third, HDBSCAN allows

Table 4
Emotion Regulation Strategy Classes Inferred With the Construct Mining Pipeline

ER strategy class	Related clusters	Description
Substances (medication)	2	Taking drugs or medication
Substances (alcohol)	13	Drinking alcohol or consuming other legal drugs
Situational control	3, 8, 15, 20, 25, 26, 27, 31	Eliminating the cause for the emotional situation
Distraction	5	Distracting from the situation
Relaxation techniques	6	Breathing techniques, walking, taking a break, etc.
Therapy	7	Seeing a professional therapist
Suppression	10	Suppressing one's feelings or one's emotional expression
Eating and drinking	12	Treating oneself with food or nonalcoholic drinks
Information seeking	17	Seeking further information about the emotional situation
Withdrawal	18	Withdrawing from the situation
Emotional expression	21, 32	Letting one's feelings run free
Instrumental support	22, 36, 37	Seeking support with a task
Self-enhancement	24	Encouraging oneself by listing positive characteristics
Social support	28, 29, 30	Seeking support from friends and family
Reappraisal	34	Seeing a situation in a positive light

Note. $N_{\text{participants}} = 113$, $N_{\text{sentences}} = 6,064$. ER = emotion regulation.

points to be left unclustered, which prevents the algorithm from unusual formulations or infrequent classes to distort the clustering. Since other clustering algorithms (e.g., k -means, agglomerative clustering, or Gaussian mixture models) violate one or more of the aforementioned assumptions, we consider them infeasible for the detection of data-driven taxonomies.

In general, we argue that informative clusterings can only be achieved, if researchers are explicit about how the affordances of the data match the assumptions of the model, since a universal definition of a “good” clustering does not exist (Hennig, 2015). Computational grounded theory has been proposed as a framework for content analysis enriched by dictionary methods and unsupervised machine learning techniques such as topic modeling (Nelson, 2020). This computer-led approach has recently been criticized because of some unwarranted assumptions (Carlsen & Ralund, 2022). One of the assumptions is that unsupervised models are natively able to locate meaningful clusters. For instance, Carlsen and Ralund (2022) demonstrated in a simulation study that latent Dirichlet allocation (LDA, a form of topic modeling or clustering) (Blei et al., 2003) is unable to uncover topics of unequal size in a constructed data set, a reasonable assumption for data in the wild. Especially in the development of artificial intelligence (AI) algorithms, engineers tend to focus on optimizing statistical metrics without a clear connection to theory (Baden et al., 2022). It has been argued for more than a decade that clustering is not an application-independent mathematical problem (von Luxburg et al., 2012). On the contrary, “good” clusterings typically include explicit descriptions of desirable cluster characteristics and adequate evaluation criteria with respect to the application at hand (Hennig, 2015).

In our example on ER, we apply the number of clusters and the unclassified samples as evaluation criteria. Although there is no agreed-upon number of ER strategy classes, previous evidence suggests a reasonable range (i.e., there might not only be two classes, but also not 100) (Southward et al., 2018). If previous evidence is unavailable, we can derive a prior from how we think about the construct. For example, if we conceptualize interoception (Vaitl, 1996), the human ability to sense bodily signals, we might theorize that there are at most as many classes of bodily sensations as there are organs in the human body. On top of that, the construct mining

pipeline does not require to specify the exact number of desired clusters. Rather, researchers can opt for a more conservative solution resulting in a larger number of clusters in earlier steps of the pipeline (Step 6), and narrow it down during evaluation and interpretation (Steps 8 and 9). For example, a more conservative solution might produce three clusters with identical interpretation (e.g., “social support”) that can be combined into one class, instead of producing a larger, potentially impure cluster in the beginning. Since we tailored the data collection such that each sentence reflects an instance of the construct of interest, we aim at reducing the number of unclassified samples as much as possible.

One way to match the specifications of the data to the assumptions of the model is to choose appropriate algorithms for existing data. Another complementary strategy is to match the type of data to the affordances of the model through tailored data collection. While data quality is a salient concern in social sciences, the AI community just starts to acknowledge the importance of data in the development and deployment of AI systems, a movement called data-centric AI (Jarrahi et al., 2023). Similarly, the construct mining pipeline defines the collection of data as its first step supporting the validity of conclusions drawn from the pipeline.

Grimmer et al. (2021) stress that human-in-the-loop approaches are essential for the evaluation of models used in exploratory research. However, judging the goodness of a model merely by interpretability or consistency with expectations is insufficient, especially since humans tend to distort evidence in favor of their attitudes and desires known as confirmation bias (Nickerson, 1998). Unsupervised machine learning models need validity measures beyond face validity or statistical fit metrics (Baden et al., 2022). With the quantification and reduction of item bias, as well as the intrusion task, the construct mining pipeline offers validity checks at multiple steps of the process that connect well to theory, while being expressed in measurable quantities eventually supporting human decision making. Future studies might also test the influence of the remaining degrees of freedom (e.g., magnitude of item bias reduction) along the pipeline empirically.

Another point of criticism for computer-led approaches (Nelson, 2020) is the misconception that researchers are able to gain substantial knowledge without immersing into the data (Carlsen & Ralund,

2022). We establish several points of contact between the researcher and the data along the pipeline. In the stage of data collection, the researcher already develops a rough understanding about the data through the design of the open-ended questions. During item bias reduction, the researcher needs to reflect on the relationship between the items and the resulting data. Finally, the researcher is encouraged to engage in deep reading during cluster interpretation.

Still, the construct mining pipeline is not without limitations. First, our method underlies the assumption that instances of the same construct class are phrased similarly over different participants and items. The data collection step provides opportunities to assert this assumption. For example, we included an exemplary item to demonstrate the response mode and provided only a limited amount of space for each ER strategy suggestion.

Second, the method only works for constructs where instances of the construct can be expressed in short texts or sentences. Even if indicators of a construct can be verbalized, participants might not be aware of all aspects of their thoughts and behaviors, or ashamed to report certain experiences. Those challenges can partly be addressed through sampling by recruiting participants knowledgeable about the construct, or with instructions at the stage of data generation. Other potential applications for the construct mining pipeline can be interoception, love languages, or counter speech. Interoception is the human ability to sense, interpret and process bodily signals (Vaitl, 1996). Similar to the case of ER, a standardized taxonomy of interoceptive dimensions does not exist, although several components have already been identified. Love languages (Chapman, 1990) describe different ways in which people express their affection. While the concept has received a lot of attention in popular culture, scientific evidence is extremely scarce (Egbert & Polk, 2006; Pett et al., 2023). Counter speech (Benesch et al., 2016) includes speech directly responding to hateful comments on the Internet. Different classifications of counter-speech strategies exist that have either been developed qualitatively (Buerger, 2021; Lasser et al., 2023) or theory-driven (Friess et al., 2021). In all of those cases, the construct mining pipeline can complement previous findings with data-driven, yet rigorous taxonomies.

The pipeline can also be used for psychological tests that require the hand coding of answers. For example, alternative application tasks (Jäger et al., 1997) assess creativity by judging the amount and originality of created ideas. Originality is assessed with a classification scheme listing classes in descending order of originality, and providing examples for common answers within each class. Our method can build classification schemes semi-automated with the possibility for regular revision and empirical measures for the prevalence of ideas. Combined with a classification model (i.e., a machine learning algorithm for automated data labeling), creativity assessment can become much faster and economic. Given that the development and scoring of tests would require less effort, test developers can design larger amounts of parallel items preventing habituation effects.

Third, we suspect that our method is more effective for purely behavioral or cognitive constructs. Current transformer models might not be able to distinguish the nuanced linguistic differences describing what somebody thinks as opposed to what somebody does. In the example on ER, we still find cognitive (e.g., reappraisal) alongside behavioral ER strategies (e.g., situational control). Results might, however, be more differentiated, if we separate the assessment of cognitive and behavioral strategies.

Fourth, the data are not entirely independent from the applied items. In Step 3, we test for the existence of item bias. We still see a leaning of answers toward each of their open-ended questions in our example after bias reduction. Nevertheless, item bias statistics allows us to quantify the impact of data collection at all, an advantage that other methods do not have. Compared to classical scale development, it is easier to think about conditions influencing the construct, incorporate them in open-ended questions, and measure their impact on the data, than to think of all possible construct dimensions itself. Concrete questions can help participants to access their understanding of the construct more easily, and influences of the data collection can be eliminated or reduced afterwards (Step 4). We believe that assessing a construct in a larger sample of laypersons has benefits over construct definitions by experts, especially in terms of range of construct dimensions. In general, the construct mining pipeline can be repeated with different data sets stemming from slightly modified instructions providing an additional robustness check. For ER strategies, we could, for example, ask to consider the effectiveness of strategies or not, whether participants would apply the strategies personally or not, or we could ask for particularly detrimental strategies.

Theory is not constituted by categorization systems alone, since they do not give any information about the relationship between variables or their predictive power. However, there is a nonnegligible, though complex connection between measurement and theory with both components informing each other (Navarro, 2021). The fact that knowledge requires good measurement tools and vice versa is termed the problem of coordination (Irvine, 2021). The social sciences prefer to specify theories in advance, a framework that forces researchers to form testable hypotheses even before collecting or viewing any data (Grimmer et al., 2021). Yet, a data-first strategy reduces the space of possible theories considerably. The universal law of generalization (Shepard, 1987), which says that the generalization over stimuli depends on their psychological distance, constitutes a successful case of measurement informing theory (Navarro, 2021). Another concern is that once the definition of a construct has become popular, the conceptualization is often taken as a given and treated as unrelated to the empirical research process (Grimmer et al., 2021). This practice might be one reason for psychological theories fading away rather than following the principle of cumulative knowledge observable in other natural sciences (Meehl, 1978). Overall, this suggests that it would be beneficial to integrate the formation of psychological constructs in the empirical research process, especially since methods keep growing and data processing costs keep lowering.

With the construct mining pipeline, we propose an alternative method for the definition of psychological constructs by combining advantages of qualitative and quantitative approaches integrating classical psychological and computational techniques. We thereby hope to provide another angle on psychological theory building by complementing top-down approaches with a data-driven, yet structured perspective.

References

- Aldao, A., Sheppes, G., & Gross, J. J. (2015). Emotion regulation flexibility. *Cognitive Therapy and Research*, 39(3), 263–278. <https://doi.org/10.1007/s10608-014-9662-4>
- Andreotta, M., Nugroho, R., Hurlstone, M. J., Boschetti, F., Farrell, S., Walker, I., & Paris, C. (2019). Analyzing social media data: A mixed-

- methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, 51(4), 1766–1781. <https://doi.org/10.3758/s13428-019-01202-8>
- Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS ONE*, 9(9), Article e106361. <https://doi.org/10.1371/journal.pone.0106361>
- Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1), 1–18. <https://doi.org/10.1080/19312458.2021.2015574>
- Balasubramanian, M., & Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, 295(5552), Article 7. <https://doi.org/10.1126/science.295.5552.7a>
- Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6), 1397–1410. <https://doi.org/10.1002/asi.23786>
- Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). Counterspeech: A literature review. *Dangerous Speech Project*. <https://dangerousspeech.org/wpcontent/uploads/2016/10/Considerations-for-Successful-Counterspeech.pdf>
- Berger, J., & Packard, G. (2022). Using natural language processing to understand people and culture. *American Psychologist*, 77(4), 525–537. <https://doi.org/10.1037/amp0000882>
- Bhati, K. S., Hoyt, W. T., & Huffman, K. L. (2014). Integration or assimilation? Locating qualitative research in psychology. *Qualitative Research in Psychology*, 11(1), 98–114. <https://doi.org/10.1080/14780887.2013.772684>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. Proceedings of the 30th International Conference on Neural Information Processing Systems (pp. 4356–4364).
- Borsboom, D., van der Maas, H. L. J., Dalege, J., Kievit, R. A., & Haig, B. D. (2021). Theory construction methodology: A practical framework for building theories in psychology. *Perspectives on Psychological Science*, 16(4), 756–766. <https://doi.org/10.1177/1745691620969647>
- Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. University of Texas at Austin. <https://www.liwc.app>
- Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, 34(5), 599–612. <https://doi.org/10.1002/per.2254>
- Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, 18, 63–68. <https://doi.org/10.1016/j.cobeha.2017.07.017>
- Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, 31(4), 340–346. <https://doi.org/10.1177/09637214221096485>
- Buerger, C. (2021). Counterspeech: A literature review. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4066882>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- Campbell, D. T., & Fiske, D. W. (1994). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. <https://doi.org/10.1037/h0046016>
- Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz,
- C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu, (Eds.), *Advances in knowledge discovery and data mining*. Lecture notes in computer science (Vol. 7819, pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14
- Carlsen, H. B., & Ralund, S. (2022). Computational grounded theory revisited: From computer-led to computer-assisted text analysis. *Big Data & Society*, 9(1), Article 205395172210801. <https://doi.org/10.1177/20539517221080146>
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22). Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf
- Chapman, G. (1990). *The 5 love languages: The secret to love that lasts*. Northfield Publishing.
- De France, K., & Hollenstein, T. (2017). Assessing emotion regulation repertoires: The regulation of emotion systems survey. *Personality and Individual Differences*, 119, 204–215. <https://doi.org/10.1016/j.paid.2017.07.018>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. ArXiv. <https://arxiv.org/abs/1810.04805>
- Di Natale, A., & Garcia, D. (2023). Lexpander: Applying colexification networks to automated lexicon expansion. *Behavior Research Methods*, 56, 952–967. <https://doi.org/10.3758/s13428-023-02063-y>
- Doré, B., Ort, L., Braverman, O., & Ochsner, K. N. (2015). Sadness shifts to anxiety over time and distance from the National Tragedy in Newtown, Connecticut. *Psychological Science*, 26(4), 363–373. <https://doi.org/10.1177/0956797614562218>
- Egbert, N., & Polk, D. (2006). Speaking the language of relational maintenance: A validity test of Chapman's (1992) five love languages. *Communication Research Reports*, 23(1), 19–26. <https://doi.org/10.1080/17464090500535822>
- Fast, E., Chen, B., & Bernstein, M. S. (2016). *Empath: Understanding topic signals in large-scale text*. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4647–4657). <https://doi.org/10.1145/2858036.2858535>
- Fine, G. A., & Elsbach, K. D. (2000). Ethnography and experiment in social psychological theory building: Tactics for integrating qualitative field data with quantitative lab data. *Journal of Experimental Social Psychology*, 36(1), 51–76. <https://doi.org/10.1006/jesp.1999.1394>
- Folkman, S., & Lazarus, R. S. (1980). An analysis of coping in a middle-aged community sample. *Journal of Health and Social Behavior*, 21(3), 219–239. <https://doi.org/10.2307/2136617>
- Ford, B. Q., Gross, J. J., & Gruber, J. (2019). Broadening our field of view: The role of emotion polyregulation. *Emotion Review*, 11(3), 197–208. <https://doi.org/10.1177/1754073919850314>
- Francis, M. E., & Pennebaker, J. W. (1992). Putting stress into words: The impact of writing on physiological, absentee, and self-reported emotional well-being measures. *American Journal of Health Promotion*, 6(4), 280–287. <https://doi.org/10.4278/0890-1171-6.4.280>
- Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, 208, 191–197. <https://doi.org/10.1016/j.jad.2016.10.019>
- Fried, E. I. (2020a). Corrigendum to “the 52 symptoms of major depression: lack of content overlap among seven common depression scales,” [Journal of Affective Disorders, 208, 191–197]. *Journal of Affective Disorders*, 260, Article 744. <https://doi.org/10.1016/j.jad.2019.05.029>
- Fried, E. I. (2020b). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4), 271–288. <https://doi.org/10.1080/1047840X.2020.1853461>

- Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, 38(5), 624–646. <https://doi.org/10.1080/10584609.2020.1830322>
- Garcia, D., & Rimé, B. (2019). Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, 30(4), 617–628. <https://doi.org/10.1177/0956797619831964>
- Garnefski, N., & Kraaij, V. (2007). The cognitive emotion regulation questionnaire. *European Journal of Psychological Assessment*, 23(3), 141–149. <https://doi.org/10.1027/1015-5759.23.3.141>
- Götz, F., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000540>
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(1), 395–419. <https://doi.org/10.1146/annurev-polisci-053119-015921>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. ArXiv. <https://arxiv.org/abs/2203.05794>
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3), 271–299. <https://doi.org/10.1037/1089-2680.2.3.271>
- Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2), 348–362. <https://doi.org/10.1037/0022-3514.85.2.348>
- Hahsler, M., Matthew, P., & Derek, D. (n.d.). *R package dbscan - density-based spatial clustering of applications with noise (DBSCAN) and related algorithms*. Retrieved July 28, 2023, from <https://cran.rproject.org/web/packages/dbscan/readme/README.html>
- Haines, S. J., Gleeson, J., Kuppens, P., Hollenstein, T., Ciarrochi, J., Labuschagne, I., Grace, C., & Koval, P. (2016). The wisdom to know the difference: Strategy-situation fit in emotion regulation in daily life is associated with well-being. *Psychological Science*, 27(12), 1651–1659. <https://doi.org/10.1177/0956797616669086>
- Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, 64, 53–62. <https://doi.org/10.1016/j.patrec.2015.04.009>
- Hilger, L., Hellwig, S., & Schulze, R. (2012, September – 27). *Deutschsprachige Adaptation des STEU sowie des STEM und erste validitätsevidenz* [German adaptation of the STEU and STEM, as well as initial validity evidence] [Paper presentation]. The 48th Congress of the German Psychological Society, Bielefeld, Germany.
- Hofmann, S. G. (2014). Interpersonal emotion regulation model of mood and anxiety disorders. *Cognitive Therapy and Research*, 38(5), 483–492. <https://doi.org/10.1007/s10608-014-9620-1>
- Hommel, B. E., Wollang, F.-J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2), 749–772. <https://doi.org/10.1007/s11336-021-09823-9>
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218. <https://doi.org/10.1007/BF01908075>
- Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, 16(4), 844–853. <https://doi.org/10.1177/1745691620970558>
- Izadpanah, S., Barnow, S., Neubauer, A. B., & Holl, J. (2019). Development and validation of the Heidelberg Form for Emotion Regulation Strategies (HFERST): Factor structure, reliability, and validity. *Assessment*, 26(5), 880–906. <https://doi.org/10.1177/1073191117720283>
- Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner intelligenzstruktur-test handanweisung* [Berlin intelligence structure test]. Hogrefe.
- Jarrah, M. H., Memariani, A., & Guha, S. (2023). The principles of data-centric AI. *Communications of the ACM*, 66(8), 84–92. <https://doi.org/10.1145/3571724>
- Jin, X., & Han, J. (2010). K-means clustering. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 563–564). Springer US. https://doi.org/10.1007/978-0-387-30164-8_425
- John, O. P., & Gross, J. J. (2004). Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. *Journal of Personality*, 72(6), 1301–1334. <https://doi.org/10.1111/j.1467-6494.2004.00298.x>
- Jolliffe, I. (2011). Principal component analysis. In M. Lovric (Ed.), *International Encyclopedia of statistical science* (pp. 1094–1096). Springer. https://doi.org/10.1007/978-3-642-04898-2_455
- Jurafsky, D., & Martin, J. H. (Eds.). (2024). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (3rd ed. Draft). <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>
- Kalinowski, T. (n.d.). *reticulate: Interface to 'Python'*. Retrieved July 28, 2023, from <https://cran.r-project.org/web/packages/reticulate/index.html>
- Kidd, S. A. (2002). The role of qualitative research in psychological journals. *Psychological Methods*, 7(1), 126–138. <https://doi.org/10.1037/1082-989X.7.1.126>
- Kjell, O. (n.d.). *text: Analyses of text using transformers models from huggingface, natural language processing and machine learning*. Retrieved July 28, 2023, from <https://cran.r-project.org/web/packages/text/index.html>
- Kjell, O. N. E., Sikström, S., Kjell, K., & Schwartz, H. A. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports*, 12(1), Article 3918. <https://doi.org/10.1038/s41598-022-07520-w>
- Kjell, O. N. E., Kjell, K., Garcia, D., & Sikström, S. (2019). Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychological Methods*, 24(1), 92–115. <https://doi.org/10.1037/met0000191>
- Kobylińska, D., & Kusev, P. (2019). Flexible emotion regulation: How situational demands and individual differences influence the effectiveness of regulatory strategies. *Frontiers in Psychology*, 10, Article 72. <https://doi.org/10.3389/fpsyg.2019.00072/full>
- Konopka, T. (n.d.). *umap: Uniform manifold approximation and projection*. Retrieved July 28, 2023, from <https://cran.r-project.org/web/packages/umap/index.html>
- Kraaij, V., & Garnefski, N. (2019). The behavioral emotion regulation questionnaire: Development, psychometric properties and relationships with emotional problems and the cognitive emotion regulation questionnaire. *Personality and Individual Differences*, 137, 56–61. <https://doi.org/10.1016/j.paid.2018.07.036>
- Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1), 61–70. <https://doi.org/10.1177/001316447003000105>
- Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: I. Hierarchical systems. *The Computer Journal*, 9(4), 373–380. <https://doi.org/10.1093/comjnl/9.4.373>
- Lasser, J., Herderich, A., Garland, J., Aroyehun, S. T., Garcia, D., & Galesic, M. (2023). *Collective moderation of hate, toxicity, and extremity in online discussion*. ArXiv. <https://doi.org/10.48550/arXiv.2303.00357>
- Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37(4), 426–441. <https://doi.org/10.1108/00483480810877598>
- MacCann, C., & Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: Theory and data. *Emotion*, 8(4), 540–551. <https://doi.org/10.1037/a0012746>
- Manning, C. D., Raghavan, P., & Schuetz, H. (2009). *An introduction to information retrieval*. Cambridge University Press.
- Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51–69. <https://doi.org/10.1080/15366367.2017.1348108>

- Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). *Mayer–Salovey–Caruso Emotional Intelligence Test (MSCEIT) users manual*. Multi-Health Systems.
- McInnes, L. (n.d.). *UMAP: Uniform manifold approximation and projection for dimension reduction*. Retrieved June 27, 2023, from <https://umap-learn.readthedocs.io/en/latest/>
- McInnes, L., Healy, J., & Astels, S. (n.d.). *The hdbscan clustering library*. Retrieved June 27, 2023, from <https://hdbscan.readthedocs.io/en/latest/>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction*. ArXiv. <https://arxiv.org/abs/1802.03426>
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Springer.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Moravcsik, A. (2014). Transparency: The revolution in qualitative research. *PS: Political Science & Politics*, 47(1), 48–53. <https://doi.org/10.1017/S1049096513001789>
- Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., & Sander, J. (2014). *Density-based clustering validation*. Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 839–847). <https://doi.org/10.1137/1.9781611973440.96>
- Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, 16(4), 707–716. <https://doi.org/10.1177/1745691620974769>
- Nelis, D., Quoidbach, J., Hansenne, M., & Mikolajczak, M. (2011). Measuring individual differences in emotion regulation: The emotion regulation profile-revised (ERP-R). *Psychologica Belgica*, 51(1), 49–91. <https://doi.org/10.5334/pb-51-1-49>
- Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42. <https://doi.org/10.1177/0049124117729703>
- Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1), 202–237. <https://doi.org/10.1177/0049124118769114>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Pellert, M., Schweighofer, S., & Garcia, D. (2021, November). Social media data in affective science. In *Handbook of computational social science* (Vol. 1, 1st ed., pp. 240–255). Routledge. <https://doi.org/10.4324/9781003024583-18>
- Pett, R. C., Lozano, P. A., & Varga, S. (2023). Revisiting the languages of love: An empirical test of the validity assumptions underlying Chapman's (2015) five love languages typology. *Communication Reports*, 36(1), 54–67. <https://doi.org/10.1080/08934215.2022.2113549>
- Pokorny, J. J., Norman, A., Zanesco, A. P., Bauer-Wu, S., Sahdra, B. K., & Saron, C. D. (2018). Network analysis for the visualization and analysis of qualitative data. *Psychological Methods*, 23(1), 169–183. <https://doi.org/10.1037/met0000129>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners* [Technical report]. OpenAI.
- Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-Networks*. <https://arxiv.org/abs/1908.10084>
- Reynolds, D. (2009). Gaussian mixture models. In S. Z. Li & A. Jain (Eds.), *Encyclopedia of biometrics* (pp. 659–663). Springer US. https://doi.org/10.1007/978-0-387-73003-5_196
- Risch, J., Möller, T., Gutsch, J., & Pietsch, M. (n.d.). *deepset/gbert-large-sts*. Deepset. Retrieved March 28, 2023, from <https://huggingface.co/deepset/gbert-large-sts>
- Rivers, S. E., Brackett, M. A., Katulak, N. A., & Salovey, P. (2007). Regulating anger and sadness: An exploration of discrete emotions in emotion regulation. *Journal of Happiness Studies*, 8(3), 393–427. <https://doi.org/10.1007/s10902-006-9017-2>
- Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, 30(2), 255–267. <https://doi.org/10.1007/s10956-020-09862-4>
- Rosenthal, R. (1994). Interpersonal expectancy effects: A 30-year perspective. *Current Directions in Psychological Science*, 3(6), 176–179. <https://doi.org/10.1111/1467-8721.ep10770698>
- Sammut, C., & Webb, G. I. (2010). TF-IDF. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (1st ed., pp. 986–987). Springer. https://doi.org/10.1007/978-0-387-30164-8_832
- Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. <https://doi.org/10.1177/1745691620966795>
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317–1323. <https://doi.org/10.1126/science.3629243>
- Sheppes, G., Scheibe, S., Suri, G., & Gross, J. J. (2011). Emotion-regulation choice. *Psychological Science*, 22(11), 1391–1396. <https://doi.org/10.1177/0956797611418350>
- Silva, V., & Tenenbaum, J. (2002). Global versus local methods in nonlinear dimensionality reduction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15). MIT Press. https://proceedings.neurips.cc/paper_files/paper/2002/file/5d6646aad9bcc0be55b2c82f69750387-Paper.pdf
- Skinner, E. A., Edge, K., Altman, J., & Sherwood, H. (2003). Searching for the structure of coping: A review and critique of category systems for classifying ways of coping. *Psychological Bulletin*, 129(2), 216–269. <https://doi.org/10.1037/0033-2909.129.2.216>
- Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, 11(1), 87–104. <https://doi.org/10.1177/1088868306294789>
- Southward, M. W., Altenburger, E. M., Moss, S. A., Cregg, D. R., & Cheavens, J. S. (2018). Flexible, yet firm: A model of healthy emotion regulation. *Journal of Social and Clinical Psychology*, 37(4), 231–251. <https://doi.org/10.1521/jscp.2018.37.4.231>
- Steck, H., Ekanadham, C., & Kallus, N. (2024). *Is cosine-similarity of embeddings really about similarity?* ArXiv. <https://arxiv.org/abs/2403.05440>
- Steinley, D. (2004). Properties of the Hubert–Arable adjusted rand index. *Psychological Methods*, 9(3), 386–396. <https://doi.org/10.1037/1082-989X.9.3.386>
- Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1), 1–34. <https://doi.org/10.1348/000711005X48266>
- Steinley, D., Brusco, M. J., & Hubert, L. (2016). The variance of the adjusted rand index. *Psychological Methods*, 21(2), 261–272. <https://doi.org/10.1037/met0000049>
- Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319–2323. <https://doi.org/10.1126/science.290.5500.2319>
- Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, 19(6), 349–357. <https://doi.org/10.1093/intqhc/mzm042>

- Troy, A. S., Shallcross, A. J., & Mauss, I. B. (2013). A person-by-situation approach to emotion regulation: Cognitive reappraisal can either help or hurt, depending on the context. *Psychological Science*, 24(12), 2505–2514. <https://doi.org/10.1177/0956797613496434>
- UKPLab. (n.d.). *Sentence transformers documentation*. Retrieved March 28, 2023, from <https://sbert.net/>
- Vaitl, D. (1996). Interoception. *Biological Psychology*, 42(1-2), 1–27. [https://doi.org/10.1016/0301-0511\(95\)05144-9](https://doi.org/10.1016/0301-0511(95)05144-9)
- van der Maaten, L., & Hinton, G. (2008). Visualizing data using T-SNE. *Journal of Machine Learning Research*, 9(86), 2579–2605. <https://jmlr.org/papers/v9/vandermaaten08a.html>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. ArXiv. <https://arxiv.org/abs/1706.03762>
- von Luxburg, U., Williamson, R. C., & Guyon, I. (2012, February). Clustering: Science or art? In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *Proceedings of ICML workshop on unsupervised and transfer learning* (Vol. 27, pp. 65–79). PMLR. <https://proceedings.mlr.press/v27/luxburg12a.html>
- Wang, Y., Tian, J., Yazar, Y., Ones, D. S., & Landers, R. N. (2022). Using natural language processing and machine learning to replace human content coders. *Psychological Methods*. Advance online publication. <https://doi.org/10.1037/met0000518>
- Weber, H., Loureiro de Assunção, V., Martin, C., Westmeyer, H., & Geisler, F. C. (2014). Reappraisal inventiveness: The ability to create different reappraisals of critical situations. *Cognition and Emotion*, 28(2), 345–360. <https://doi.org/10.1080/02699931.2013.832152>
- Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., & Liu, S. (2021). Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 529–539. <https://doi.org/10.1109/TVCG.2021.3114694>
- Zepeda-Mendoza, M. L., & Resendis-Antonio, O. (2013). Hierarchical agglomerative clustering. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of systems biology* (pp. 886–887). Springer. https://doi.org/10.1007/978-1-4419-9863-7_1371

Appendix A

Lexicon Extension Methods

Lexicons, or word lists, are a common method to retrieve psychological concepts from text. One of the most widely known collections of lexicons in psychology is the Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022), which was first developed in a study on language and disclosure (Francis & Pennebaker, 1992). Each lexicon comprises a list of words related to a certain concept (e.g., analytical thinking, but also emotions such as anger, anxiety, or sadness), which can be counted in a text. The higher the relative word count, the stronger the representation of a concept in a text.

Lexicons come with certain disadvantages. For example, words can have different meanings or connotations in different contexts (e.g., in legal vs. everyday speech). Words can also change in meaning or connotation over time. Additionally, new words might emerge that are then commonly used to describe an existing construct or state (e.g., the development of emojis with the rise of the Internet).

Therefore, the adaptation of lexicons might be necessary to accommodate different use cases.

Lexicon extension methods are a means to extend existing word lists or to create new lexicons based on a small number of seed words. There are multiple approaches to lexicon extensions with and without relation to linguistic theory. One of the first approaches to expand word lists is based on collections of synonyms (WordNet) (Miller, 1995). A recent approach grounded in linguistic theory uses colexification networks for lexicon expansion (LEXpander) (Di Natale & Garcia, 2023). Colexification occurs if a single word expresses two different, but related concepts (e.g., Greek uses the same word for medicine and poison). A third, computational approach includes finding related words to a word list based on nearest neighbors in word embeddings (Empath) (Fast et al., 2016).

For interested readers, we refer to the above-mentioned literature.

Appendix B

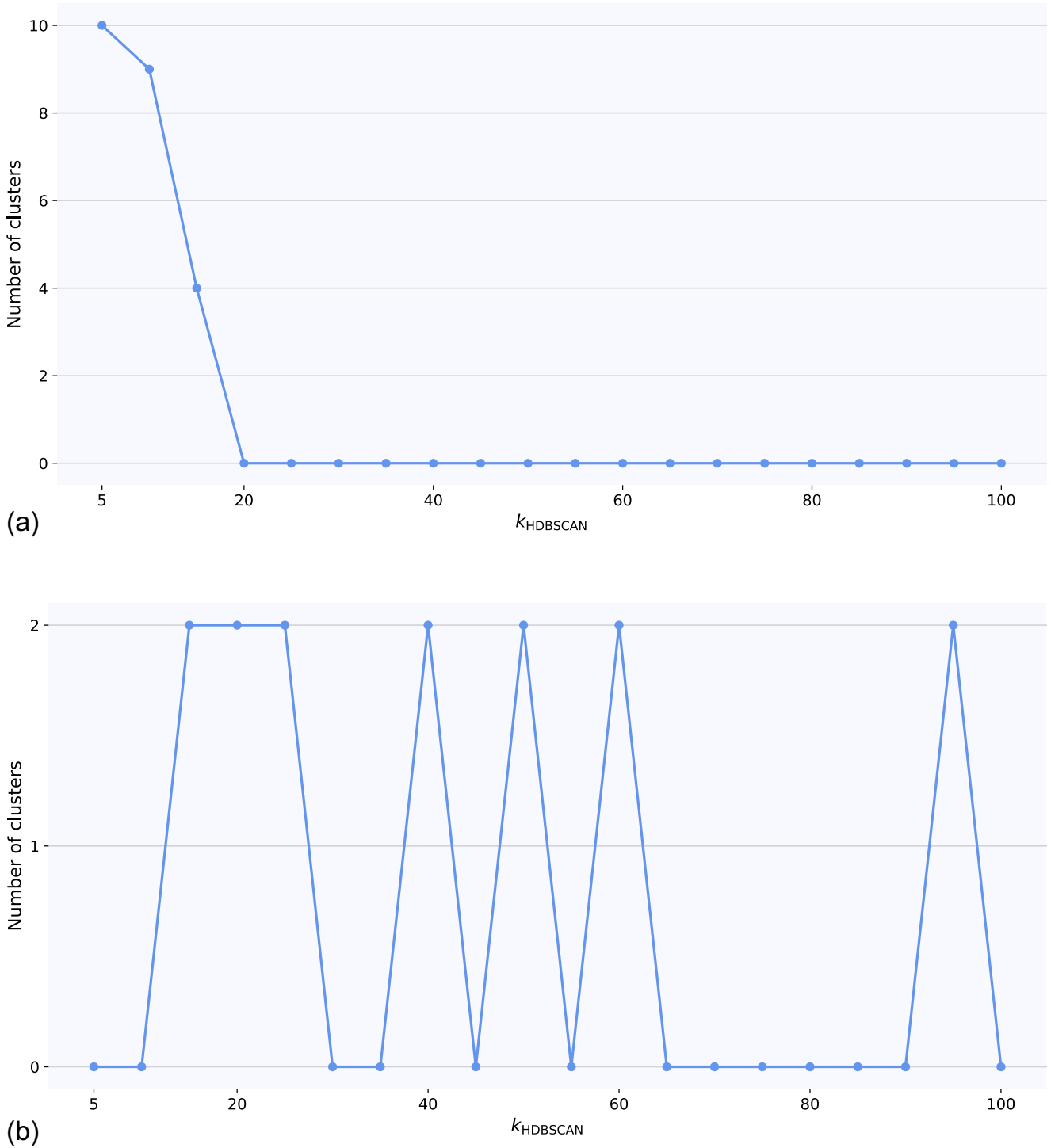
The Curse of Dimensionality

The curse of dimensionality refers to the phenomenon that data becomes sparse in high-dimensional spaces and therefore loses meaning. In high dimensions, points are all approximately equally far away from each other. As a result, when clustering all points

likely fall into one large cluster. We illustrate the curse of dimensionality in Figure B1, where we applied HDBSCAN to the original sentence embeddings of 1,024 dimensions without UMAP reduction.

Figure B1

Clustering on Sentence Embeddings Without Dimension Reduction Applying HDBSCAN With (a) Euclidean Distance and (b) Cosine Distance



Note. k_{HDBSCAN} = number of nearest neighbors for distance calculation in HDBSCAN. For clustering with the Euclidean distance, the number of discovered clusters quickly drops to zero (i.e., all samples are classified as noise) when increasing the number of nearest neighbors. For cosine distance, the number of discovered clusters oscillates between zero and two. HDBSCAN = hierarchical density-based spatial clustering of applications with noise. See the online article for the color version of this figure.

(Appendices continue)

Appendix C

Emotion Regulation Strategy Inventiveness Task

In the following, we provide full instructions, an exemplary item, and all nine emotional vignettes of the Emotion Regulation Strategy Inventiveness Task (ERSIT) in English (translated) and German (original).

Instructions

English (Translated)

The following test was designed to measure creativity when dealing with emotional situations.

The test consists of nine subtasks. At the beginning of each subtask you will see a short description of an everyday situation that either elicits fear, anger, or sadness. Read the description carefully. Try to imagine the given situation as vividly as possible. Your task will be to list as many ideas as possible, for changing one's own feelings for the better in the short or the long run; independent of whether you would choose to carry out the strategies in the end or not. The listed ideas should be preferably distinct from each other. Let your thoughts run free and list everything that comes to your mind.

Sufficient extra time will be assigned for reading each situation description. The time for listing your ideas will be limited to 3 min. In the end, please additionally select the idea that is the most effective in your opinion. The investigator will watch the time and lead you

through the test. On the following page, you will be presented with an exemplary item, which will illustrate the task.

German (Original)

Mit dem folgenden Test möchten wir erfassen, wie groß dein Einfallsreichtum im Umgang mit emotionalen Situationen ist.

Der Test besteht aus 9 Aufgaben. Zu Beginn jeder Aufgabe liest du eine kurze Beschreibung einer Alltagssituation, die entweder Angst, Ärger oder Traurigkeit auslöst. Lies die Situationsbeschreibungen jeweils aufmerksam durch. Versuche dich möglichst lebhaft in die beschriebene Situation hineinzusetzen. Deine Aufgabe wird es sein, möglichst viele Ideen aufzulisten, wie man das eigene Befinden kurz- oder langfristig zum Besseren wenden kann; unabhängig davon, ob du diese Ideen am Ende selbst umsetzen würdest. Es ist umso besser, wenn du möglichst viele verschiedene Ideen auflistest. Lass deinen Gedanken freien Lauf und notiere alles, was dir einfällt.

Für das Lesen der Situationsbeschreibungen steht dir jeweils ausreichend Zeit zur Verfügung. Die Zeit, in der du deine Antworten notieren kannst, ist auf 3 Minuten beschränkt. Markiere im Anschluss bitte zusätzlich die in deinen Augen wirkungsvollste Idee. Der/Die Versuchsleiter/in wird auf die Zeit achten und dich durch den Test führen. Auf der folgenden Seite siehst du zunächst ein Musterbeispiel, das dir die Aufgabe verdeutlichen wird.

Exemplary Item

Figure C1

Exemplary ERSIT Item in English (Translated)

Please vividly imagine the following situation:

A good colleague and you have shared an office for years, but your colleague gets a new job and you lose contact with her. You are sad.

Please list all ideas for changing your own feelings for the better in the short or the long run that you can think of.

✕ I call her regularly.

2 I make friends with my new colleague.

3 I invite my old colleague for lunch.

4 I just forget my old colleague.

5 I eat a large sundae.

6 I just apply for the same job as my old colleague.

7 I meet my old colleague for coffee.

8 _____

9 _____

10 _____

...

Note. ERSIT = emotion regulation strategy inventiveness task.

Figure C2

Exemplary ERSIT Item in German (Original)

Bitte versetze dich möglichst lebhaft in die folgende Situation:

Eine gute Kollegin und du haben sich seit Jahren ein Büro geteilt. Deine Kollegin bekommt einen neuen Job und du verlierst den Kontakt zu ihr. Du bist traurig.

Notiere alle Ideen, die dir einfallen, wie man das eigene Befinden in der Situation kurz- oder langfristig zum Besseren wenden kann.

✕ Ich rufe sie regelmäßig an und frage, wie es so läuft.

2 Ich freunde mich mit meiner neuen Kollegin an.

3 Ich lade meine alte Kollegin zum Essen ein.

4 Ich vergesse meine alte Kollegin einfach.

5 Ich esse einen großen Eisbecher.

6 Ich bewerbe mich einfach für die selbe Abteilung wie meine alte Kollegin.

7 Ich treffe mich mit meiner alten Kollegin in einem Kaffee.

8 _____

9 _____

10 _____

...

Note. ERSIT = emotion regulation strategy inventiveness task.

(Appendices continue)

Emotional Vignettes

Table C1

Emotional Vignettes and Respective Masked Words in English (Translated)

Vignette number	Vignette text	Masked words
1	Your workmate fails to deliver an important piece of information on time, causing you to fall behind schedule also	Information, schedule
2	You are accepted for a highly sought after contract, but have to fly to the location. You have a phobia of flying	Contract, company, phobia of flying
3	You answer the phone and hear that close relatives are in hospital critically ill	Hospital
4	You find out that some member of your social sports team have been saying that you are not a very good player	Sports team, team, <i>member</i>
5	You have just gone back to university after a lapse of several years. You are surrounded by younger students who seem very confident about their ability and you are unsure whether you can compete with them	Lapse, university, <i>student</i>
6	A demanding client takes up a lot of your time and then asks to speak to your boss about your performance. Although your boss assures you that your performance is fine, you feel upset	Client
7	Your access to essential resources has been delayed and your work is way behind schedule. Your progress report makes no mention of the lack of resources	<i>Resource</i> , schedule, progress report, progress
8	You are having a large family gathering to celebrate you moving into your new home. You want the day to go smoothly and are a little nervous about it	Large family, gathering, celebrate
9	You and your colleague usually go to a cafe after the working week and chat about what's going on in the company. After your colleague's job is moved to a different section in the company, he/she stops coming to the cafe. You miss these Friday talks	Cafe, section

Note. Italic words have been modified with regular expressions to account for inflections and declensions. See analysis code for details.

Table C2

Emotional Vignettes and Respective Masked Words in German (Original)

Vignette number	Vignette text	Masked words
1	Dein Kollege kann eine wichtige Information nicht rechtzeitig bereitstellen, sodass auch du deinen Zeitplan nicht einhalten kannst	Information, Zeitplan
2	Du hast einen heiß begehrten Vertrag angeboten bekommen. Das Unternehmen ist an einem weit entfernten Ort ansässig und du musst dorthin fliegen. Du leidest jedoch unter Flugangst	Vertrag, Unternehmen, Flugangst
3	Du gehst ans Telefon und erfährst, dass nahe Verwandte schwer krank im Krankenhaus liegen.	Krankenhaus
4	Du findest heraus, dass einige Mitspieler/innen deines Sportteams gesagt haben, du seist kein/e gute/r Spieler/in	Sportteam, Team, <i>Mitspieler</i>
5	Nach einigen Jahren Auszeit bist du gerade wieder an die Universität zurückgekehrt. Du bist umgeben von jüngeren Student/innen, die sehr überzeugt von ihren Fähigkeiten scheinen und du bist dir nicht sicher, ob du mit ihnen mithalten kannst	Auszeit, Universität, <i>Student</i>
6	Ein schwieriger Klient nimmt viel von deiner Zeit in Anspruch. Aufgrund deiner Leistung möchte er sich anschließend bei deinem Chef beschweren. Obwohl dein Chef dir versichert, dass deine Leistung gut sei, bist du niedergeschlagen	Klient
7	Dein Zugang zu wichtigen Betriebsmitteln hat sich hinausgezögert und dein Arbeitsfortschritt ist deswegen weit hinter dem Zeitplan. In deinem Zwischenbericht wird das fehlende Betriebsmittel nicht erwähnt	<i>Betriebsmittel</i> Zeitplan, Zwischenbericht, Arbeitsfortschritt
8	Deine Großfamilie kommt zusammen, um deinen Einzug in deine neue Wohnung zu feiern. Du möchtest, dass der Tag problemlos verläuft und bist deswegen nervös	Großfamilie, Feier, feiern
9	Du und ein gut befreundeter Kollege treffen sich normalerweise am Ende der Woche in einem Café und unterhaltet euch über die Vorkommnisse im Unternehmen. Nachdem die Stelle deines Kollegen in eine andere Abteilung verlegt wurde, kommt er nicht mehr in das Café. Du vermisst die Unterhaltung am Freitag	Café, Cafe, Abteilung

Note. Italic words have been modified with regular expressions to account for inflections and declensions. See analysis code for details.

(Appendices continue)

Appendix D

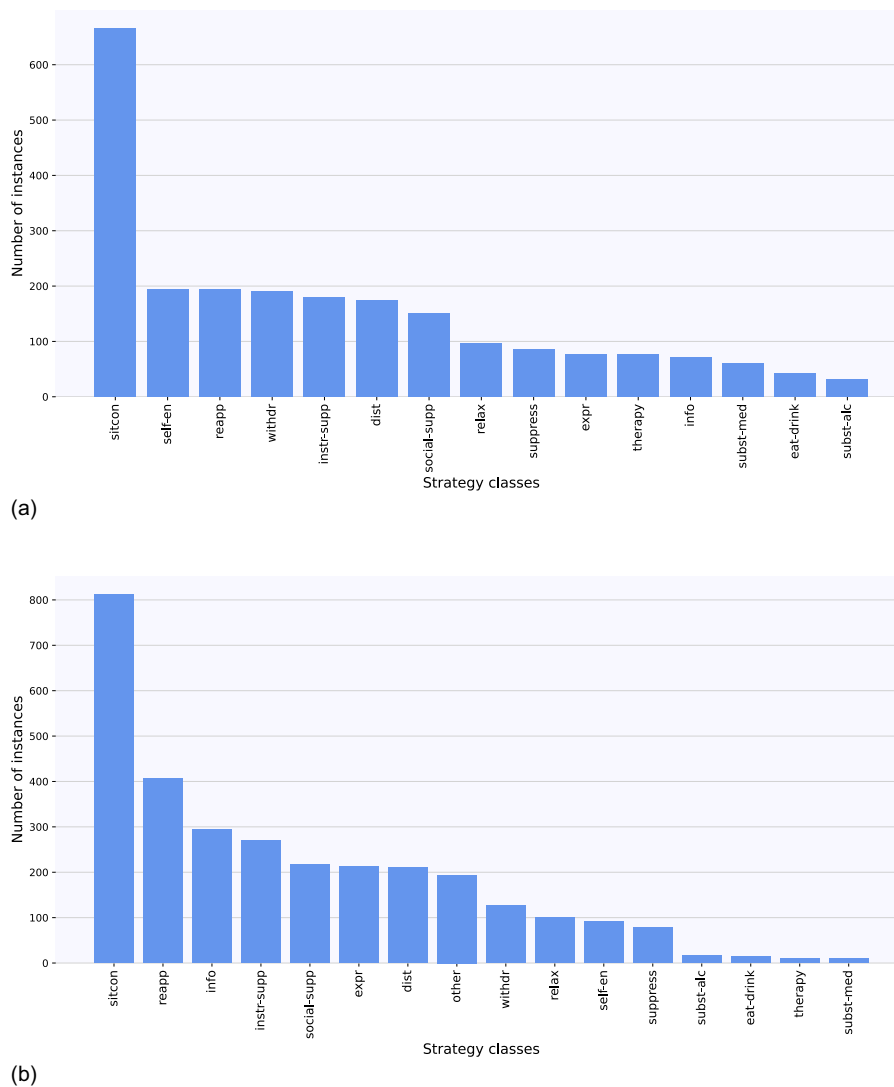
Analysis of Unclassified Samples

Since there is a substantial amount of unclassified samples after applying UMAP and HDBSCAN (Steps 5 and 6), we investigated their content via human annotation. If the construct mining pipeline is able to uncover the structure of psychological constructs by identifying high-density regions in the semantic space, we should be able to classify the remaining samples with the discovered classification scheme (Step 9), and find no to very few additional classes. Instead, we hypothesize that samples will fall out of clusters due to their wording rather than their content.

We hired one undergraduate and two postgraduate psychology students as research assistants to complete the annotation task. We provided them with a description of the discovered classification system (see Table 4 of the main article). The document contained a definition of emotion regulation strategies, a guiding question and principles for annotation, as well as descriptions of all classes including their title, abbreviation, and an example. We provided the research assistants with disjoint sets of unclassified sentences of equal size ($n = 1,022$ for two and $n = 1,023$ for one of the annotators), and asked them to

Figure D1

Label Distribution in (a) Classified Samples (Samples in Clusters) and (b) Unclassified Samples



Note. sitcon = situational control, self-en = self-enhancement, reapp = reappraisal, withdr = withdrawal, instr-supp = instrumental support, dist = distraction, relax = relaxation techniques, social-supp = social support, suppress = suppression, expr = emotional expression, therapy = therapy, info = information seeking, subst-med = substances (medication), social-supp = social support, eat-drink = eating and drinking, subst-alc = substances (alcohol). See the online article for the color version of this figure.

(Appendices continue)

label the samples with said classification system. We added the class “other” and told them to assign this label whenever they felt that a sentence does not fall into any of the provided classes. Furthermore, we asked them to propose a new label for sentences, for which they decided to assign the class “other.” The process was guided by a kick-off meeting, regular check-ins and a closing meeting.

In order to estimate interrater reliability, we further provided each of the annotators with a subsample of the other annotator’s batches, that is, two subsamples of $n = 205$ amounting to 20% of one batch. We used Krippendorff’s α (Krippendorff, 1970) to calculate interrater reliability, a measure that is suitable for imbalanced classification tasks with multiple raters. Krippendorff’s α was $\alpha = .45$, $\alpha = .49$, and $\alpha = .51$ for each of the cross-labeled subsamples. We suspect that the moderate reliability can be traced back to sentences, where multiple classes apply, since we cannot know the intention of the writer. For example, in the situation where members of one’s sports team have been saying that one is not a good player, some people suggest to “approach the members about the issue.” This suggestion can either be classified as information seeking (e.g., investigating whether or why the members have been saying this), situational control (e.g., trying to make the members stop those claims), or even emotional expression (e.g., telling the members how those claims made you feel). However, for our investigation it is not important to know in which class a sentence falls exactly, but rather that it falls into any of the discovered classes at all.

Figure D1 shows the distribution of labels within samples that fell into a cluster (classified samples), and samples that did not (unclassified samples). The position of labels within the distribution is roughly the same for both sets. For example, situational control is by far the dominating class and reappraisal comes shortly after. The most infre-

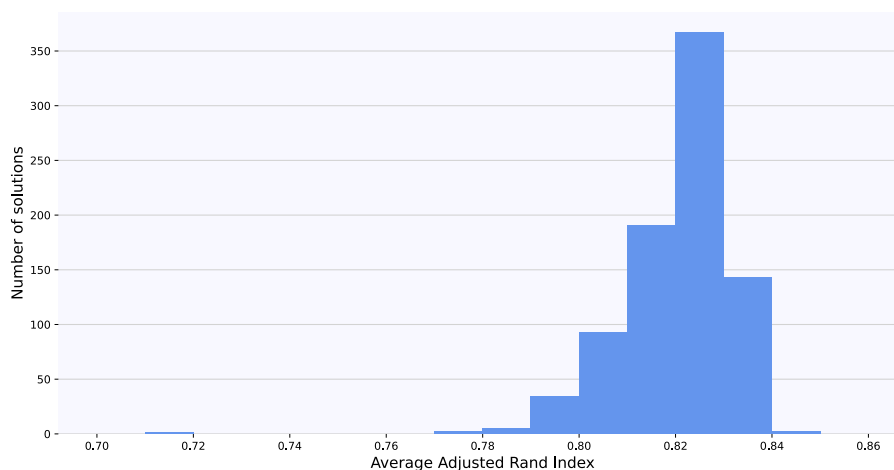
quent classes comprise substances (alcohol and medication), as well as eating and drinking. Information seeking is much more frequent among unclassified samples, while withdrawal and self-enhancement are more infrequent. Classes might have different levels of dependence on open-ended questions. For instance, information seeking will most often refer to the content of the situation (e.g., “I ask A why B”), while withdrawal and self-enhancement can be expressed in more general terms (e.g., “I quit” or “I tell myself I can do it”).

Overall, the annotators labeled 194 sentences in the unclassified samples as “other,” that is, 6.3% of all unclassified samples and 3.2% of the entire data set. The annotators agreed on four new classes within “other”: acceptance ($n = 76$), reflection ($n = 73$), destructive social behavior ($n = 42$), and self-deprecation ($n = 3$). Acceptance refers to accepting the situation without valuation (e.g., “I accept the situation and continue”). Reflection means to ponder about the incident without ruminating (e.g., “I ask myself why I am sad”). Destructive social behavior includes aggressive verbal and physical behavior, or thoughts and actions of revenge (e.g., “I call them names”). Self-deprecation means blaming oneself for negative events (e.g., “I blame myself for not taking enough care of that person”). Reflection is close to information seeking. While information seeking aims to clarify a situation by involving others, reflection is looking for explanations within oneself. As discussed in the main article, cognitive strategies might be hard to detect with the current setup, and separate surveys for cognitive and behavioral strategies might be sensible. Likewise, current instructions focus on strategies that participants consider being adaptive. A separate survey might be necessary to uncover rather maladaptive strategies, such as destructive social behavior or self-deprecation. We conclude that occurrences of undetected classes are sufficiently small and can be traced back to suboptimal data collection procedures.

Appendix E

Distribution of Average Adjusted Rand Indices

Figure E1
Average Rand Indices for Different Random Seeds



Note. Average rand indices calculated after excluding trivial solutions (i.e., number of clusters = 2). See the online article for the color version of this figure.

(Appendices continue)

Appendix F

Cluster Top Words

Table F1

Top 10 Characteristic Words for Each Cluster Normalized by the Frequency of Words Within and Across Clusters (German Original)

Cluster	Intrusion score	Top words
1	0.34	darüber, für, über, könnte, reden, freunden, rede, können, hören, Arbeit
2	0.96	Medikamente, nehme, Beruhigungstabletten, Beruhigungsmittel, nehmen, Tabletten, besorge, beruhigende, Flug, Beruhigungstropfen
3	0.96	Zug, Alternativen, suche, Verkehrsmittel, Schiff, Bahn, fahren, Transportmöglichkeiten, Ort, suchen
4	0.94	fliegen, Flug, fliege, Angst, Statistiken, trotzdem, Person, Sicherheit, informiere, einfach
5	0.79	ablenken, lenke, ab, abzulenken, Musik, Ablenkung, Film, Sport, schauen, schau
6	0.97	entspannen, meditiere, Pause, Entspannungstechniken, Entspannungsübungen, tief, atmen, mache, atme, Atemübungen
7	0.91	Therapie, Flugangst, Therapeuten, gehe, gehen, überwinden, Psychologen, psychologische, therapieren, versuchen
8	0.88	besuchen, Krankenhaus, besuche, Verwandten, fahre, Verwandte, Besuch, fahren, sofort, Person
9	0.65	Zeit, Spaß, erledigen, sinnvoll, Dinge, macht, derweil, tun, Hobbies, Sachen
10	0.86	ignorieren, vergessen, ignorieren, versuche, vergesse, versuchen, weitermachen, verdrängen, darüber, Gedanken
11	0.75	rufe, anrufen, Kontakt, telefonieren, telefoniere, Freundin, Kollegen, halten, beste, regelmäßig
12	1.00	essen, esse, Schokolade, Eis, gutes, Kaffee, gehen, trinken, kochen, gut
13	0.94	trinke, Alkohol, Glas, betrinke, trinken, Wein, vorher, weinen, Entspannung, Bier
14	0.96	anfreunden, Freunde, Kollegen, Freundschaften, neue, Freunden, befreunden, treffe, anzufreunden, versuche
15	0.95	treffen, Termin, Tag, Kollegen, frage, ausmachen, finden, trotzdem, möchte, neuen
16	0.75	Café, gehen, gehe, freitags, alleine, allein, Kaffee, Kollegen, mehr, einladen
17	0.91	Krankheit, informieren, informiere, erkundige, Behandlungsmöglichkeiten, Arzt, Heilungschancen, recherchieren, krank, beschäftigen
18	0.88	ab, wechseln, wechsele, Job, Team, lehne, Vertrag, Studium, suchen, suche
19	0.85	Angst, Ängste, spreche, Familie, Freunden, sprechen, Ängste, reden, Nervosität, Freundinnen
20	0.89	bereite, plane, Tag, gut, planen, vorbereiten, do, to, Liste, Plan
21	1.00	Trauer, weinen, weine, lasse, freien, Lauf, Traurigkeit, Gefühle, verkrieche, Gefühlen
22	0.85	Hilfe, Rat, bitte, helfen, fragen, Unterstützung, Freunde, bitte, Kollegen, holen
23	0.44	Zeit, Geschenk, schöne, Verwandten, Person, verbringen, erleben, Momente, Erinnerungen, Zeiten
24	0.76	Fähigkeiten, stärken, eigenen, Augen, bewusst, gut, mehr, Erfahrung, Erfolge, eigene
25	0.83	härter, trainiere, mehr, verbessern, trainieren, besser, anstrengen, versuche, beim, Leistung
26	0.88	lerne, lernen, mehr, Lerngruppen, Nachhilfe, Lerngruppe, Vorlesungen, bilden, Tutorien, beitreten
27	1.00	Zeitplan, Zwischenbericht, überarbeiten, anpassen, ändern, korrigieren, neuen, neu, optimieren, umstrukturieren
28	0.85	reden, darüber, Freunden, jemandem, jemanden, sprechen, unterhalten, drüber, auslassen, jemand
29	1.00	darüber, rede, spreche, Freunden, Freundin, Freund, erzähle, davon, gute, Kollegin
30	0.96	Familie, reden, Freunden, darüber, sprechen, Verwandten, fremden, vertrauten, Person, anrufen
31	0.81	Information, beschaffen, Informationen, selber, bekommen, versuchen, Info, mal, versuche, wichtige
32	0.94	Ärger, lasse, Wut, Frust, rauslassen, Sport, Luft, beim, boxen, rauszulassen
33	1.00	beschwere, beschweren, beim, Chef, Kollegen, Freunden, zuständigen, darüber, Vorgesetzten, aufregen
34	0.79	schlimm, Schuld, gut, daran, sage, bewusst, sagen, perfekt, klar, schon
35	0.80	Bericht, Beschwerde, zuständigen, Grund, Person, frage, Klienten, wer, weise, Klient
36	0.90	Chef, Situation, Vorgesetzten, erkläre, Sache, darüber, spreche, Vorgesetzte, aufzuklären, berichte
37	1.00	Chef, Trainer, rede, ansprechen, darauf, aussprechen, Aussprache, erneut, besprechen, Bescheid

Note. Clusters 1, 9, and 23 fall below the majority threshold in the intrusion task (Step 8). Clusters 4 (“fliegen”), 11 (“anrufen”), 14 (“anfreunden”), 16 (“Café”), 19 (“Angst”), and 33 (“beschweren”) formed based on topic words. Cluster 35 was not interpretable regardless of a high score in the intrusion task.

Received August 9, 2023
Revision received April 1, 2024
Accepted June 18, 2024 ■