Discriminant Validity: Disentangling Health and Emotional Constructs from Language-Based Assessments

Scott Feltman[1], **Adithya V. Ganesan**[1,2], **Whitney Ringwald**[3], **Roman Kotov**[1], **Benjamin J. Luft**[1], **Ryan L. Boyd**[4], **H. Andrew Schwartz**[1,2], **Oscar Kjell**[2,5]
[1]Stony Brook University, [2]Vanderbilt University, [3]University of Minnesota Twin Cities, [4]University of Texas at Dallas, [5]Lund University
scott.feltman@stonybrook.edu, nils.e.kjell@vanderbilt.edu

Abstract

Language-based assessments of health have demonstrated validity in terms of convergence with self-report questionnaire scores, but recent work has shown they lack an important psychometric property: discriminant validity, the ability of the measure to distinguish the target construct from a related one. For example, a language-based general mental health assessment may accurately assess the degree of mental health, but it may also produce scores that correlate with general physical health well beyond theoretical expectations. Here, we propose and evaluate two methods for altering the standard loss function to directly penalize off-target correlations while retaining convergent validity. Evaluated across two clinical language datasets spanning physical and mental health, the augmentations substantially reduced off-target correlations with minimal loss in convergent validity (on-target correlations). Limiting loss in convergent validity to 0.005 Pearson correlation points, the loss function using Squared Cosine Similarity Discrimination significantly improved discriminant validity of language-based mental and physical health assessments ($r = 0.454 \rightarrow r = 0.322$; $p < 0.05$); and fundamental psychopathology dimensions ($r = 0.442 \rightarrow 0.430$; $p < 0.05$). These findings demonstrate that enforcing discriminant validity as a training objective is effective, moving language-based assessments closer to the specificity required for differential clinical use.

1 Introduction

Language provides a rich signal for understanding mental and physical health (Boyd and Schwartz, 2020; Tausczik and Pennebaker, 2010; Kjell et al., 2023). Computational psycholinguistic research has demonstrated valid language-based assessments of depression (Eichstaedt et al., 2018), anxiety (Rutowski et al., 2020), well-being (Kjell et al., 2021; Mesquiti et al., 2026), and physical health outcomes (Lian et al., 2023; Adejumo et al., 2024). Language sources have ranged from natural social media text (Pak and Paroubek, 2010; Merchant et al., 2019; Alvi et al., 2023) to probed responses, where individuals describe their psychological state in open-ended natural language in response to targeted questions (Kjell et al., 2018) to structured clinical interviews (Hur et al., 2024). With the advent of pretrained large language models, the accuracy (i.e., the convergent validity) of probed language-based assessments has improved to approach the theoretical upper limit of accuracy defined by the target scales' own reliability with $r > 0.80$ (Kjell et al., 2021; Gu et al., 2025; Nilsson et al., 2024).

However, a fundamental psychometric challenge remains understudied in the language-based assessment literature: discriminant—validity the degree to which a measure is unrelated to conceptually distinct constructs (Campbell and Fiske, 1959). Mental and physical health constructs are rarely independent; when language-based models are trained to predict one construct, they frequently yield inflated correlations with correlated off-target constructs (Kjell et al., 2021; Gu et al., 2025), making it difficult to determine whether a model is capturing something specific to the target condition or merely reflecting shared variance across constructs.

Prior work has largely focused on maximizing convergent validity without penalizing correlations with off-target constructs. While some approaches within the language-based assessment literature have partially addressed this through measurement design (Kjell et al., 2018, 2021; Gu et al., 2025; Zhang et al., 2018; Wu et al., 2020), discriminant validity has not been enforced as a penalty within the training objective for continuous health outcome regression. In this work, we address this gap by augmenting the ridge regression objective function with two discriminant penalty terms, deriving closed-form solutions for each, and evaluating their

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 43–55

behavior across two health language datasets with differing construct structures. Specifically, we address the following research questions:

- To what degree can objective function augmentation in ridge regression enforce discriminant validity between correlated health constructs assessed from language while retaining convergent validity?
- To what degree does objective function augmentation affect the inter-correlation of predictions from models trained on distinct constructs?
- Does discriminant improvement across constructs systematically vary according to the original (non-penalized) convergent validity or the target scale's reliability?

2 Background

Discriminant Validity in Health Assessment. Discriminant validity, introduced by Campbell and Fiske (1959) as part of the multitrait-multimethod framework, refers to the degree to which a measure does not spuriously correlate with conceptually distinct constructs. Discriminant validity is a cornerstone of psychometric assessment (Campbell and Fiske, 1959; Phan et al., 2016; Meirte et al., 2016). Valid measurement requires not only that a measure converges with other measures of the same construct, but that it diverges from measures of different constructs—a dual requirement that maps directly onto the challenge of language-based health assessments. A model trained to assess depression from language should correlate strongly with depression rating scales, but substantially less with anxiety or somatic scales. A measure that cannot make this distinction holds limited interpretive value regardless of its convergent accuracy (Campbell and Fiske, 1959; Jackson, 1969; Marsh and Bailey, 2016; Meyer et al., 2001).

This challenge is particularly pronounced in mental health assessment, where comorbidity is common (Kessler et al., 2005; Krueger and Markon, 2006). Mental health constructs frequently co-occur, such as depression with anxiety, substance use disorders with antisocial behavior, or narcissism with paranoid personality disorder (Ruggero et al., 2019; Kotov et al., 2017). Mental health conditions also commonly co-occur with physical health problems (Prince et al., 2007; Scott et al., 2008; DE Hert et al., 2011), creating overlapping construct spaces that challenge discriminant measurement across both domains. In clinical settings, the ability to distinguish between related conditions is essential for accurate diagnosis (Regier et al., 2013; Gu et al., 2025), treatment planning (Gu et al., 2015; Insel et al., 2010), and outcome monitoring (Hunsley and Mash, 2007), making discriminant validity not merely a psychometric requirement but a practical necessity for meaningful health assessment.

Discriminant Validity in Language-Based Assessment. Even with construct-specific probed responses, such as individuals separately describing their level of each construct in open-ended natural language (Kjell et al., 2018), language-based assessments show poor to moderate discriminant validity; predicted score correlations between harmony in life and satisfaction with life reach $r = 0.96$, despite the rating scales themselves correlating at $r = 0.85$ (Kjell et al., 2021).

The problem compounds when language sources are combined across constructs and response formats: convergent validity increases, but assessment scores become nearly indistinguishable, with cross-construct correlations reaching $r = 0.97$ (Gu et al., 2025). The problem of poor discriminant validity has also been found across language sources, including Fried (2016), Sikström et al. (2024b), and Rasing et al. (2017). Models trained on such data risk learning average expressions of related constructs rather than the specific signal of each, reducing the interpretive precision of both predictions and the linguistic features associated with them. Prior work has addressed this through measurement design by assessing *normalized difference scores* between constructs (Kjell et al., 2021; Gu et al., 2025) or using cosine similarity between individuals' responses to construct-specific word norms (Kjell et al., 2018, 2021). This demonstrates that a construct-specific signal exists in language and can be partially recovered; however, none of these approaches use poor discriminant validity as a penalty within the training objective itself. Rather, they either redefine the prediction target (i.e., training to a difference score) or address it at the representation level (i.e., using similarity to construct-specific word norms), leaving discriminant validity as a design choice rather than an optimization target.

Discriminant Loss Functions in Natural Language Processing (NLP). Penalizing representational overlap during training has been explored in

NLP for related but distinct purposes: contrastive learning has been used to separate dissimilar representations in self-supervised settings (Chen et al., 2020; Wu et al., 2020; Giorgi et al., 2021), and adversarial training has been applied to remove demographic confounds from text representations (Zhang et al., 2018; Elazar and Goldberg, 2018). Most relevant to our approach, discriminant ridge classifiers have been proposed for categorical outcomes, penalizing overlap between class predictions to improve sensitivity and specificity (Peng and Cheng, 2021). However, extending this logic to continuous regression is non-trivial: categorical discriminant penalties operate on class boundaries and discrete decision regions, whereas continuous health outcomes require penalizing the correlation between real-valued prediction vectors—a fundamentally different geometric objective. None of these approaches has therefore been adapted for continuous health outcome regression, where the goal is to reduce the correlation between predictions of related yet distinct continuous constructs. This gap motivates the present work.

Contributions. We propose two augmentations to the ridge equation: *Mean Squared Error Discrimination* (MSE Discrimination) and *Squared Cosine Similarity Discrimination* (SCS Discrimination) that enforce discriminant validity between correlated health constructs assessed from language while retaining convergent validity. We derive closed-form solutions for both augmentations compatible with standard Singular Value Decomposition (SVD)-based solvers (Eckart and Young, 1936) and evaluate how objective function augmentation affects the discriminant validity of language-based assessment scores from models trained on distinct constructs. Next, we show that discriminant validity improvements systematically vary as a function of the scale's baseline convergent validity and reliability. We also introduce a preliminary investigation into the disentanglement of linguistic features resulting from enforcing discriminant validity, demonstrating the potential to extract specific linguistic features and partially remove associative strength from opposing, but related constructs. Together, these contributions address a practical barrier to the clinical deployment of language-based health assessment: without discriminant validity, models risk conflating related conditions, limiting their utility for differential assessment, the identification of construct-specific linguistic markers, and the development of targeted, individualized treatment plans.

3 Datasets

We evaluate our methods on two datasets: the *World Trade Center (WTC) Clinic* dataset, consisting of transcripts from automated clinical interviews with physical and mental health quality of life outcomes (Kjell et al., 2026), and the *Interviews of Hierarchical Taxonomy of Psychopathology (iHiTOP)* dataset, comprising transcripts from semi-structured clinical interviews and responses to the iHiTOP questionnaire with six psychopathology spectrum scores (see Kotov et al. (2024) for more details). The Institutional Review Board approved both studies at *Stony Brook University* (IRB#2022-00391 and IRB#604113).

3.1 WTC Clinic

Participants were enrolled through the *Stony Brook WTC Health and Wellness Program*, a clinical program providing ongoing health monitoring to responders to the September 11th attacks (Kjell et al., 2026). During yearly clinic visits, participants completed structured screening questionnaires covering physical and mental health symptoms. The final dataset contained transcripts from N = 1,589 participants, with 849 words on average. The cohort was overwhelmingly male (91.93% male, 7.40% female) and predominantly white (84.81% Caucasian, 3.22% African American, 0.56% Asian American, 7.78% multi-racial). Summary statistics of the WTC Clinic dataset are provided in Table 1. Outcomes were the Physical Component Summary (PCS) and Mental Component Summary (MCS) scores derived from the SF-12 (Steenstrup et al., 2013), measuring quality of life across physical and mental health dimensions, respectively.

3.2 iHiTOP

The Hierarchical Taxonomy of Psychopathology (HiTOP), introduced by (Kotov et al., 2017), organizes psychopathological characteristics into a dimensional hierarchy based on empirical patterns of co-variation. Despite being comprised of distinct items (Kotov et al., 2017), individual scales share high correlations as seen in Table 2, making it a particularly demanding testbed for discriminant validity for psychopathology as a whole. Participants (N=609) completed a survey comprising over 200 items spanning the HiTOP framework. Six

WTC Clinic	Number of Participants	Average	SD
Participants	1,589		
Tokens	-	848.92	633.44
Age	-	57.93	8.00
PCS	1,308	46.07	10.64
MCS	1,308	52.49	8.86
iHiTOP	**Number of Participants**	**Average**	**SD**
Participants	609		
Tokens	-	9,055.46	6,159.76
Age	-	43.49	16.11
Internalizing	598	0.69	0.49
Somatoform	574	0.28	0.46
Thought Disorder	596	0.27	0.39
Detachment	596	0.52	0.48
Disinhibition	605	0.64	0.56
Antagonism	603	0.33	0.36

Table 1: Descriptive statistics for WTC Clinic and iHiTOP datasets, detailing language contents for analyzed transcripts with physical and mental health constructs.

	INT	SOM	TD	DET	DIS	ANT
INT	1.000	0.394	0.450	0.660	0.653	0.482
SOM	0.394	1.000	0.293	0.281	0.267	0.243
TD	0.450	0.293	1.000	0.394	0.424	0.383
DET	0.660	0.281	0.394	1.000	0.476	0.428
DIS	0.653	0.267	0.424	0.476	1.000	0.581
ANT	0.482	0.243	0.383	0.428	0.581	1.000

Table 2: Symmetric correlation matrix of iHiTOP spectra for valid participants. High inter-scale correlation suggests low discriminant validity, leading to ambiguous feature learning during training without careful preparation.

HiTOP spectrum scores were derived as composite scores from their constituent items, requiring a minimum of approximately 80% valid item responses and otherwise treated as missing: Internalizing, Somatoform, Thought Disorder, Detachment, Disinhibition, and Antagonism. The sample was predominantly female (62.8%) and white (78.17%), with a mean age of 43.49 years (*SD* = 16.11). Descriptive statistics are provided in Table 1.

4 Methods

4.1 Closed-form solutions

To enforce discriminant validity directly within the training objective, we augment the standard Ridge Regression (Hoerl and Kennard, 1970) loss function with two discriminant penalty terms and derive closed-form solutions for each. Both solutions are compatible with standard Singular Value Decomposition (SVD)-based solvers (Eckart and Young, 1936) as implemented in scikit-learn (Pedregosa et al., 2011), enabling straightforward integration into existing language-based assessment pipelines. We derive each solution by setting the derivative of the augmented objective function to zero, following the same algebraic approach used in standard ridge regression.

MSE Discrimination. For a single pair of on- and off-target outcomes, let $\mathbf{X}$, y, d, w, α and γ be the input feature matrix containing user language representations, outcome array of the target clinical construct, outcome array of off-target clinical constructs, model weight vector which describes the relationship between language features and clinical outcomes, L2 penalty which reduces overfitting by penalizing excessive large-magnitude weights, and discriminant penalty which penalizes high accuracies between the model predictions and off-target labels, respectively. The augmented ridge equation is given as

$$L = (\mathbf{X}w - y)^2 + \alpha||w||_2^2 - \gamma(\mathbf{X}w - d)^2 \quad (1)$$

We take the derivative with respect to w and set the expression equal to zero, yielding

$$0 = 2\mathbf{X}^T(\mathbf{X}w^* - y) + 2\alpha w^* - 2\gamma\mathbf{X}^T(\mathbf{X}w^* - d) \quad (2)$$

The weight vector, now denoted by w^*, is the set of model parameters that minimizes the penalized objective function. This achieves a balance between reducing error with respect to the construct of interest and reducing predictive overlap with constructs on which the models are not being trained. By distributing the first and third terms, we can group like terms together and isolate w^* to obtain

$$\mathbf{X}^T(y - \gamma d) = [\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I} - \gamma\mathbf{X}^T\mathbf{X}]w^* \quad (3)$$

$$= [(1 - \gamma)\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I}]w^* \quad (4)$$

Factoring out $1-\gamma$ from the right-hand side, letting $G = 1 - \gamma$, and taking the inverse results in the optimal weight vector

$$w^* = \frac{1}{G}[\mathbf{X}^T\mathbf{X} + \frac{\alpha}{G}\mathbf{I}]^{-1}\mathbf{X}^T(y - \gamma d) \quad (5)$$

In this form, it is clear to see that the discriminant parameter γ behaves nonlinearly. This is in contrast to the L2 penalty, which behaves linearly within the differential equation. As a consequence, this objective function is only valid for $0 \leq \gamma < 1$[1], lest the function become non-convex—no longer guaranteeing an optimal solution—which would result in incoherent model results. Further, it is reasonable to suspect that as γ approaches 1, the convergence becomes increasingly unstable, leading to uncertainties in the solution.

It is common for datasets to have more than two variables of interest; accordingly, we can further expand the above solution for multiple discriminant dimensions as

$$w^* = \frac{1}{G}[\mathbf{X}^T\mathbf{X} + \frac{\alpha}{G}\mathbf{I}]^{-1}\mathbf{X}^T(y - \gamma d_{tot}) \quad (6)$$

where $G = 1 - N\gamma$, and N, d_{tot} are the number and sum of discriminant dimensions, respectively. It is important to note that as the number of constructs against which to discriminate increases, the maximum value of γ decreases, limiting the valid range and requiring a finer grid search net. Thus, the domain of γ can be expanded to

$$\{\gamma \in \mathbb{R} \mid 0 \leq \gamma < \frac{1}{N}\} \quad (7)$$

From Equation 7, we can conclude that MSE Discrimination is suitable for two-outcome modeling in order to keep the range of γ consistent.

Many datasets, including iHiTOP, have more than two correlated outcomes, requiring a different loss function.

Squared Cosine Similarity Discrimination. While MSE Discrimination separates related constructs by their absolute distance, this may not completely remove overlapping latent features during model fitting. Instead, it may be more beneficial to separate their *alignment*, which can be determined by cosine similarity. Cosine similarity has seen much use in the field of computational linguistics, especially with its easy-to-calculate gradient when using normalized vectors. Upon scaling outcome and prediction vectors to a magnitude of 1, the formula for cosine similarity simplifies to the dot product

$$cos(\theta) = (\mathbf{X}w) \cdot d = d^T\mathbf{X}w \quad (8)$$

This allows for a clean, convex, and numerically stable closed-form solution of the augmented ridge equation. However, because cosine similarity spans a range of [-1,1], we opted to use Squared Cosine Similarity (SCS) to prioritize minimizing the magnitude of construct similarity, instead of producing highly anti-correlated predictions. This results in the modified objective function

$$L = (\mathbf{X}w - y)^2 + \alpha||w||_2^2 - \gamma(d^T\mathbf{X}w)^2 \quad (9)$$

Following the same approach as with MSE Discrimination, we can arrive at the minimum loss solution

$$w^* = [\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I} + \gamma\mathbf{X}^T dd^T\mathbf{X}]^{-1}\mathbf{X}^T y \quad (10)$$

In contrast to the MSE discrimination penalty, γ mimics the linear behavior of α within the objective function, while removing the limited range and allowing for a more robust and well-behaved hyperparameter. However, this alteration results in a non-symmetric matrix, which adds an extra step to the SVD solver since now diagonalization cannot be exploited.

5 Results

5.1 Discriminant and Convergent Validity Under Objective Function Augmentation

Using SCS Discrimination shows a steady decline in convergent validity for SF-12 constructs, while their respective discriminant correlations decreased significantly faster and to a larger degree Figure 1. Discriminant and convergent validity stabilized for γ > 1000 at a tolerance of 1e-4, suggesting no further changes from continuing to raise gamma. MSE Discrimination demonstrated significantly more pronounced changes in both convergent and discriminant validity, still improving discrimination between SF-12 constructs until γ = 0.75; however, this augmentation becomes unstable where $0.9 < \gamma < 1$. Detailed plots of model performance

[1]While $\gamma < 0$ is physically possible, this would penalize discrimination between the off-target construct, violating its purpose.

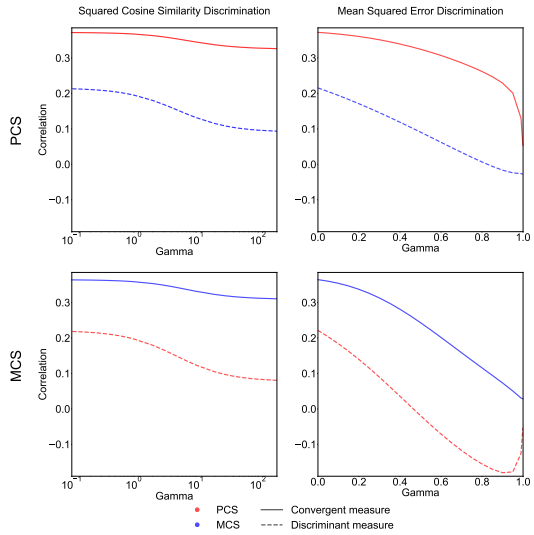

Figure 1: Performance of MSE and SCS Discrimination on SF-12 constructs. PCS and MCS show very minimal losses in convergent validity for SCS Discrimination, with pronounced decreases in off-target correlations as γ increases. With MSE Discrimination, MCS experiences a more severe loss in convergent validity relative to PCS, however divergence between off-target metrics continues to increase until $\gamma = 0.75$. The model begins to collapse at large γ, as numerical instability overpowers the closed-form solution. MSE Discrimination also loses convergent validity at a faster rate than SCS Discrimination, suggesting preferential usage of the latter augment.

for SF-12 constructs are shown in Figure 1. Because of the instability and limited range of MSE Discrimination, especially with multiple scales, only SCS Discrimination is tested on iHiTOP data.

All HiTOP constructs showed varying degrees of improvements in separation, with four out of six showing large divergence on average. Internalizing and Disinhibition show marginal increases in discrimination against off-target constructs on average, while Somatoform shows the greatest increase in separation and least loss in convergent validity. Antagonism, Detachment, and Thought Disorder display modest decreases in convergent validity with moderately increased discriminant validity for the five off-target constructs. Interestingly, four of the six constructs, without discriminant penalties, had worse convergent validity than at least one of the off-target constructs, which is quickly alleviated as gamma increases. Figure 2 shows the convergent and discriminant validity plots for all six spectra for the iHiTOP dataset, including individual constructs.

Squared Cosine Similarity Discrimination		Mean Squared Error Discrimination	
Gamma	Correlation	Gamma	Correlation
0	0.454	0	0.454
0.05	0.446	0.01	0.438
0.25	0.417	0.05	0.372
1.0	0.322	0.10	0.284
5.0	0.052	0.25	-0.015
25.0	-0.198	0.50	-0.535
∞	-0.316	1.0	-1[2]

Table 3: Inter-prediction correlations for SF12 constructs compared between separation method. For MSE Discrimination, inter-correlation begins to drop sharply after $\gamma = 0.01$, and approaches perfect anti-correlation as γ approaches 1. For squared cosine similarity, inter-correlations decline more slowly as compared to MSE Discrimination, in addition to reaching a minimum of -0.316 at a tolerance of 1e-4, preventing instability.

5.2 Discriminant Validity of Language-Based Assessments

For SF-12 constructs, the inter-correlations between language-based assessment predictions consistently decreased for both MSE and SCS Discrimination as γ increased. This suggests that the discriminant penalty does incentivize decoupling predictions from the distinct models, and that models are learning features more specific to that construct. See Table 3 for a more detailed layout of the distribution of SF-12 language-based assessment predictions.

HiTOP constructs exhibit results similar to SF-12, with language-based assessment inter-correlations strictly decreasing as γ increased, showing consistent and stable behavior with SCS Discrimination. We found that $\gamma = 5.0$ yielded the smallest magnitudes on average (Table 4), further confirmed with a Mann-Whitney U test against remaining γ values ($p < 0.05$).

5.3 Role of Scale Reliability and Baseline Convergent Validity in Discriminant Improvement

The relative improvement of discriminant validity was different for individual constructs, calling into question the existence of a systemic relationship between performance gain (divergence), and their latent properties. Because gamma selection is dependent upon the individual needs for analysis, we opt to select the largest gamma with a loss of con-

[2]Because $\{\gamma \in \mathbb{R} \mid 0 \leq \gamma < 1\}$ for MSE Discrimination, this value represents the limit as γ approaches 1.0.

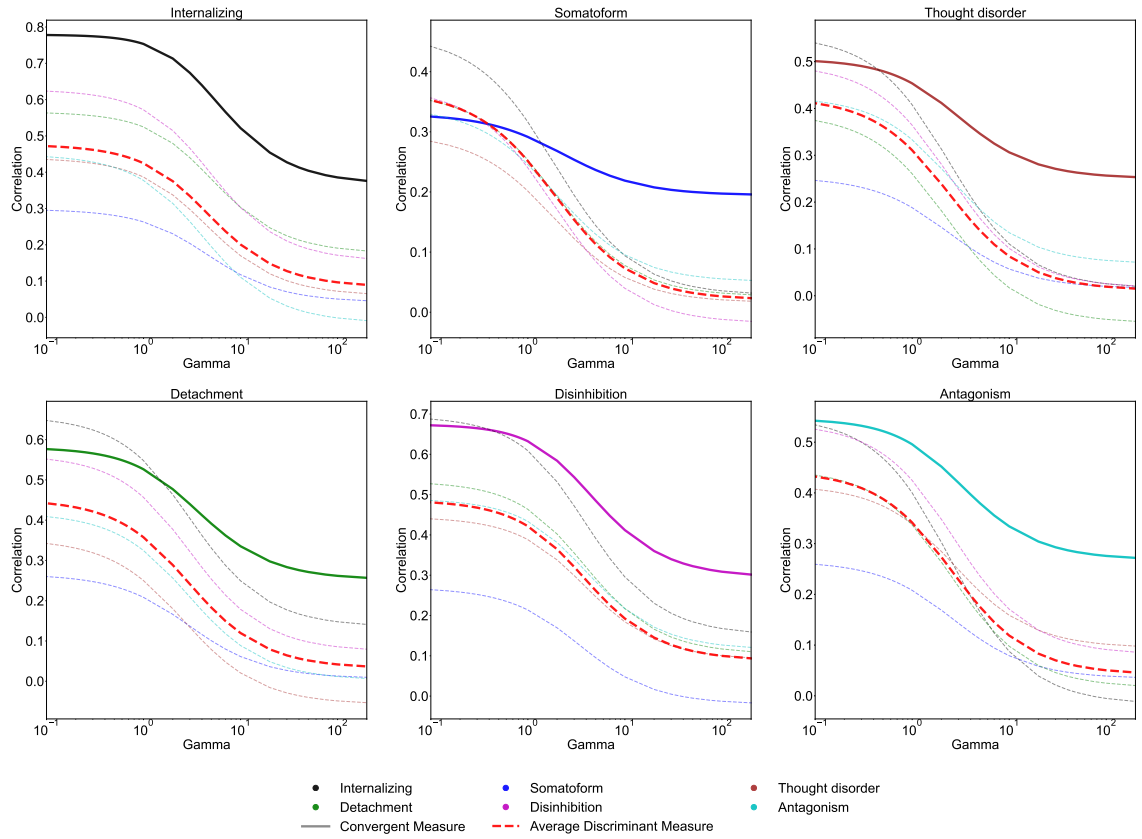

Figure 2: Performance plots of models trained on individual HiTOP constructs against average off-target correlations as a function of γ. Somatoform, Thought Disorder, and Antagonism show rapid improvements in discriminant validity. Internalizing and Detachment display modest acceleration of discriminant validity, while Disinhibition shows little improvement.

vergent validity below 0.005—a practical heuristic representing a negligible reduction that falls within the margin of sampling variability at these sample sizes and would be largely indistinguishable from the non-penalized estimate in practice. We also select the gamma at which inter-prediction correlations are closest to orthogonal as a comparison to demonstrate the varying behaviors for different strengths of discriminant penalties ($\gamma = 5.0$, as determined in subsection 5.2).

We found that all six scales yielded statistically significant improvements in discriminant validity when limiting loss in convergent validity to 0.005 ($p < 0.05$) using Benjamini-Hochberg Correction for multiple tests (Benjamini and Hochberg, 1995). Four out of six scales attained statistically significant improvements in discriminant validity at $\gamma = 5.0$ (see Table 5 for more details.) Discriminant validity gain relative to convergent validity loss at $\gamma = 5.0$ correlated extremely well with non-penalized convergent validity ($r = -0.967$, $p < 0.05$), and moderately well with the reliability coefficient for each scale ($r = -0.768$, $p < 0.10$). However, when limiting convergent validity loss to 0.005, reliability was significantly correlated with discriminant validity improvement ($r = -0.823$, $p < 0.05$).

5.4 Linguistic Feature Disentanglement

Preliminary differential language analysis was performed on the SF-12 constructs to evaluate relationships between group-level linguistic features and model predictions using the *topics* (Ackermann et al., 2024) package in R (R Core Team, 2025). Correlations were obtained for non-penalized and penalized ridge model predictions, and evaluated according to association strength change for PCS and MCS, respectively, against 1-grams in participant transcripts. Words commonly associated with physical health (such as "pain" or "breathing") were disentangled from MCS with reductions in correlations larger than 0.005, while strengthening correlations of words more associated with mental health, for both negative and positive contexts (e.g.

Spectra	Gamma	INT	SOM	TD	DET	DIS	ANT
INT	0	-	0.609	0.685	0.846	0.855	0.690
	0.30	-	0.515	0.583	0.786	0.792	0.585
	5.00	-	0.015	-0.051	0.330	0.219	-0.128
	∞	-	-0.162	-0.266	0.144	-0.060	-0.397
SOM	0	0.609	-	0.519	0.526	0.521	0.480
	0.10	0.576	-	0.483	0.487	0.479	0.440
	5.00	0.015	-	0.023	-0.088	-0.203	-0.087
	∞	-0.162	-	-0.065	-0.222	-0.393	-0.196
TD	0	0.685	0.519	-	0.587	0.671	0.653
	0.05	0.667	0.501	-	0.565	0.653	0.636
	5.00	-0.051	0.023	-	-0.181	-0.037	0.123
	∞	-0.266	-0.065	-	-0.346	-0.217	0.024
DET	0	0.846	0.526	0.587	-	0.764	0.639
	0.10	0.826	0.487	0.544	-	0.733	0.600
	5.00	0.330	-0.088	-0.181	-	0.030	-0.100
	∞	0.144	-0.222	-0.346	-	-0.206	-0.277
DIS	0	0.855	0.521	0.671	0.764	-	0.769
	0.15	0.824	0.460	0.618	0.718	-	0.731
	5.00	0.219	-0.203	-0.037	0.030	-	0.230
	∞	-0.060	-0.393	-0.217	-0.206	-	0.078
ANT	0	0.690	0.480	0.653	0.639	0.769	-
	0.10	0.654	0.440	0.620	0.600	0.743	-
	5.00	-0.128	-0.087	0.123	-0.100	0.230	-
	∞	-0.397	-0.196	0.024	-0.277	0.078	-

Table 4: Inter-prediction correlations for HiTOP constructs at chosen γ. Zero and asymptotic strengths are selected for display to show the overall dynamics of discriminant prediction validity as a function of γ. Inter-correlations consistently decrease as penalty strength increases, even into inverse relationships despite the strictly positive discriminant augment. For $\gamma = 5.0$, the average correlation among predictions was 6.32×10^{-4}.

		$\gamma = 5.0$		Optimal γ		
Spectra	Reliability	ΔDisc.	$\Delta\Delta$Div.	Gamma	ΔDisc.	$\Delta\Delta$Div.
INT	0.960	0.202	0.033	0.30	0.010	**0.005**
SOM	0.850	0.255	**0.162**	0.10	0.014	**0.010**
TD	0.844	0.286	**0.125**	0.05	0.020	**0.016**
DET	0.866	0.270	**0.079**	0.10	0.009	**0.005**
DIS	0.919	0.225	0.035	0.15	0.007	**0.003**
ANT	0.807	0.269	**0.101**	0.10	0.010	**0.019**

Table 5: Changes in discriminant validity and divergence, defined as change between loss in discriminant and convergent correlation, for six iHiTOP spectra at selected gamma. Optimal gammas were selected to limit loss of convergent validity to 0.005. Four out of six HiTOP constructs demonstrated significantly increased negative shifts in correlation at $\gamma = 5.0$ ($p < 0.05$, indicated by bold), while all iHiTOP constructs attained significant increases in discriminant validity at the optimal strengths. This suggests successful balancing of limited loss in convergent validity while strengthening the signal compared to off-target constructs.

"divorce", "grandkids"). Similar patterns were observed with PCS, decreasing correlations of language features tied to emotions like "family" or "ptsd" while strengthening associations for phrases related to physical health (see Table 6).

6 Discussion

The results demonstrate that discriminant validity can be directly enforced as an optimization target in continuous health outcome regression, addressing a gap that has persisted in the language-based assessment literature despite growing recognition of the problem (Kjell et al., 2021; Gu et al., 2025; Sikström et al., 2024a). Both augmentations reduced off-target correlations while retaining convergent validity and eventually achieving orthogonality of inter-prediction correlations, suggesting that penalization reshapes weights to produce construct-specific solutions rather than the shared variance.

The two augmentations enforced discriminant validity through distinct mechanisms, with important practical consequences. MSE Discrimination penalizes the absolute distance between predictions and the off-target outcome vector, producing sharp reductions in off-target correlations but within a constrained valid range that shrinks as additional constructs are added. The solution becomes numerically unstable as γ approaches 1, and inter-prediction correlations quickly continue into strong anti-correlation, tightening the relationship between predictions in the opposite direction, which is equally undesirable.

SCS Discrimination instead penalizes the angular alignment between prediction vectors, useful in situations where producing predictions which do not share the same associative strength as their constructs is ideal. This leads to more gradual, stable improvements that converge to a well-behaved minimum. This angular framing has conceptual precedent in contrastive learning approaches that leverage cosine similarity to separate dissimilar representations (Gao et al., 2021), and in adversarial training used to remove demographic confounds from text representations (Zhang et al., 2018; Elazar and Goldberg, 2018). However, those approaches operate on discrete classes or binary confounds, whereas the present work penalizes correlation between real-valued continuous prediction vectors—a geometrically distinct objective requiring a different closed-form treatment. MSE Discrimination may therefore be appropriate when only two constructs are involved and aggressive separation is needed, but SCS is preferable in the more common case of multiple correlated outcomes, where stability and a wide usable range of γ are necessary.

The strong inverse relationship between baseline convergent validity and discriminant gain ($r = -0.967$, $p < 0.05$) at $\gamma = 5.0$ suggests an interpretable mechanism: models of con-

Outcome	Word	Baseline Pearson R	Penalized Pearson R
PCS	family	0.161	0.153
	ptsd	-0.104	-0.096
	surgery	-0.171	-0.176
	recovered	-0.051	-0.059
MCS	pain	-0.102	-0.092
	breathing	-0.095	-0.085
	grandkids	0.062	0.069
	divorce	-0.125	-0.131

Table 6: Preliminary behavior of 1-grams correlated with SF-12 constructs before and after discriminant penalization. 1-grams from participant transcripts were correlated with predictions from standard and discriminant ridge models (selecting gamma to limit convergent validity loss to 0.005). Disentanglement is seen for both PCS and MCS, showing changes larger than the degree of convergent validity loss.

structs with high-performing language-based analysis have weight vectors robust to shifts from discriminant penalization. Constructs with lower baseline convergent validity exhibit more moldable weight vectors, allowing for effective reshaping towards construct-specific features. The reliability finding adds a complementary layer: less reliable scales, which present noisier prediction targets, benefit most in absolute discriminant terms ($r = -0.899$, $p < 0.05$) but show a weaker relationship against discriminant gain ($r = -0.767$, $p < 0.10$). Less reliable scales, however, statistically benefit to a greater degree upon limiting convergent validity loss ($r = -0.823$, $p < 0.05$), hinting at an interesting dynamic between different gammas across health dimensions. Together, these results suggest that the practical benefit of objective function augmentation is greatest precisely in the cases where language-based assessment is most challenging—low-reliability, harder-to-predict constructs—and may offer diminishing returns as convergent validity approaches its theoretical ceiling, as seen in Table 5.

The present findings may have direct implications for the clinical deployment of language-based health assessment. While the current implementation utilizes deterministic solutions, this procedure can be applied to gradient-based models, such as transformers, as current models trained on correlated health outcomes risk conflating related conditions, limiting their utility for differential assessment of health conditions (e.g., Somatoform symptoms from Internalizing disorder, or physical from mental health decline.) This further shows potential usage as a fine-tuning procedure to tailor models to specific health constructs before clinical deployment. By enforcing discriminant validity as an optimization target, models produce predictions that are more interpretable, separable, and suitable for the identification of construct-specific linguistic markers. Linguistic features show potential for disentanglement as a result of enforcing discriminant validity, with positive examples of expected directional shifts of 1-grams with SF-12 predictions (see Table 6). Though the degree of separation was small, hovering at and just above the 0.005 convergent loss limit, greater linguistic separation may emerge by using larger gamma values, at the risk of more loss in convergent validity. As language-based assessments move closer to clinical deployment (e.g. in tele-health, automated interviews, and longitudinal monitoring), the ability to distinguish between related conditions and their corresponding language, rather than merely predict them in aggregate, becomes not a psychometric nicety but a practical prerequisite for meaningful use.

7 Conclusion

We proposed and evaluated two loss functions that enforce discriminant validity as a direct optimization target in continuous health outcome regression. Both augmentations substantially improved discriminant validity while retaining convergent validity. Notably, discriminant improvement was greatest for constructs with lower baseline convergent validity and reliability. We also demonstrated an introductory procedure to evaluate the effect of discriminant validity enforcement on linguistic features, showing that enforcing discriminant validity through loss functions has the potential to disentangle words from opposing, but correlated constructs. This opens the possibility of characterizing how distinct health conditions are expressed differently in language, rather than merely predicting them in aggregate, thereby informing the theoretical understanding of what is linguistically specific to each construct. As language-based assessments move closer to clinical deployment in areas such as tele-health, automated interviews, and longitudinal monitoring, the ability to distinguish between related conditions becomes a practical prerequisite for meaningful use. The present work provides a tractable, theoretically grounded path toward that goal.

Limitations

Both solutions require closed-form derivations, which constrains the class of penalty functions that can be applied; gradient-based extensions are possible but were not evaluated here. The selection of γ requires a grid search, and the valid range for MSE Discrimination narrows with each additional construct, making hyperparameter tuning increasingly costly in high-dimensional construct spaces. The present evaluation was conducted on two datasets with specific demographic compositions and language elicitation formats; generalization to other populations, language sources, and outcome structures remains to be established. Further, to our knowledge, no other studies have explored the variability of discriminant validity in this fashion, providing no ground truth to the degree of discrimination for comparison. This is amplified by the distinct nature and large volume of psychometric constructs. Future work should examine the predictive overlap between different psychometric constructs for future comparisons. Proper evaluation of methods to assess discriminant validity should include on- and off-target correlations for each model per construct of interest, using the methods established in this paper in addition to novel frameworks. Additionally, it should be seen whether these augmentations extend to transformer-based fine-tuning pipelines and whether the construct-level moderators identified here, being baseline validity and reliability, replicate across domains beyond psychopathology as well as mental and physical health.

References

Leon Ackermann, Zhuojun Gu, and Oscar N. E. Kjell. 2024. An R-package for visualizing text in topics.

Philip Adejumo, Phyllis M Thangaraj, Lovedeep Singh Dhingra, Arya Aminorroaya, Xinyu Zhou, Cynthia Brandt, Hua Xu, Harlan M Krumholz, and Rohan Khera. 2024. Natural language processing of clinical documentation to assess functional status in patients with heart failure. *JAMA Netw Open*, 7(11):e2443925.

Quratulain Alvi, Syed Ali, Sheikh Ahmed, Nadeem Khan, Mazhar Awan, and Haitham Nobanee. 2023. On the frontiers of twitter data and sentiment analysis in election prediction: a review. *PeerJ Computer Science*, 9:e1517.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Ryan L Boyd and H Andrew Schwartz. 2020. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *J Lang Soc Psychol*, 40(1):21–41.

D T Campbell and D W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull*, 56(2):81–105.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *Preprint*, arXiv:2002.05709.

Marc DE Hert, Christoph U Correll, Julio Bobes, Marcelo Cetkovich-Bakmas, Dan Cohen, Itsuo Asai, Johan Detraux, Shiv Gautam, Hans-Jurgen Möller, David M Ndetei, John W Newcomer, Richard Uwakwe, and Stefan Leucht. 2011. Physical illness in patients with severe mental disorders. I. prevalence, impact of medications and disparities in health care. *World Psychiatry*, 10(1):52–77.

Carl Eckart and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proc Natl Acad Sci U S A*, 115(44):11203–11208.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Eiko I Fried. 2016. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *J Affect Disord*, 208:191–197.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online. Association for Computational Linguistics.

Jenny Gu, Clara Strauss, Rod Bond, and Kate Cavanagh. 2015. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? a systematic review and meta-analysis of mediation studies. *Clinical Psychology Review*, 37:1–12.

Zhuojun Gu, Katarina Kjell, H Andrew Schwartz, and Oscar Kjell. 2025. Natural language response formats for assessing depression and worry with large language models: A sequential evaluation with model pre-registration. *Assessment*, page 10731911251364022.

Arthur E. Hoerl and Robert W. Kennard. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

John Hunsley and Eric J Mash. 2007. Evidence-based assessment. *Annu Rev Clin Psychol*, 3:29–51.

Jihyun K Hur, Joseph Heffner, Gloria W Feng, Jutta Joormann, and Robb B Rutledge. 2024. Language sentiment predicts changes in depressive symptoms. *Proc Natl Acad Sci U S A*, 121(39):e2321321121.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry*, 167(7):748–751.

Douglas N. Jackson. 1969. Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, 72(1):30–49.

Ronald C Kessler, Wai Tat Chiu, Olga Demler, Kathleen R Merikangas, and Ellen E Walters. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Arch Gen Psychiatry*, 62(6):617–627.

Oscar Kjell, Daiva Daukantaitė, and Sverker Sikström. 2021. Computational language assessments of harmony in life - not satisfaction with life or rating scales - correlate with cooperative behaviors. *Front Psychol*, 12:601679.

Oscar N E Kjell, Scott Feltman, H. A Schwartz, Adithya V Ganesan, Whitney R Ringwald, Sean Clouston, Melissa A Carr, Benjamin Luft, and Roman Kotov. 2026. Distinguishing the language of mental and physical health: A sequential evaluation with model preregistration of automated clinical visit interviews.

Oscar N E Kjell, Katarina Kjell, Danilo Garcia, and Sverker Sikström. 2018. Semantic measures: Using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol Methods*, 24(1):92–115.

Oscar N E Kjell, Katarina Kjell, and H Andrew Schwartz. 2023. Beyond rating scales: With targeted evaluation, large language models are poised for psychological assessment. *Psychiatry Res*, 333:115667.

Roman Kotov, Robert F Krueger, David Watson, Thomas M Achenbach, Robert R Althoff, R Michael Bagby, Timothy A Brown, William T Carpenter, Avshalom Caspi, Lee Anna Clark, Nicholas R Eaton, Miriam K Forbes, Kelsie T Forbush, David Goldberg, Deborah Hasin, Steven E Hyman, Masha Y Ivanova, Donald R Lynam, Kristian Markon, and 21 others. 2017. The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J Abnorm Psychol*, 126(4):454–477.

Roman Kotov, Holly Levin-Aspenson, Katherine Jonas, Camilo Ruggero, and Others. 2024. Interview for the Hierarchical Taxonomy of Psychopathology (iHiTOP). `https://osf.io/u25em/`.

Robert F Krueger and Kristian E Markon. 2006. Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. *Annu Rev Clin Psychol*, 2:111–133.

Ruixue Lian, Vivian Hsiao, Juwon Hwang, Yue Ou, Sarah E Robbins, Nadine P Connor, Cameron L Macdonald, Rebecca S Sippel, William A Sethares, and David F Schneider. 2023. Predicting health-related quality of life change using natural language processing in thyroid cancer. *Intell Based Med*, 7.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Herbert Marsh and Michael Bailey. 2016. Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis. *The Journal of Higher Education*, 64:1–18.

Jill Meirte, Ulrike Van Daele, Koen Maertens, Peter Moortgat, Rudi Deleus, and Nancy E Van Loey. 2016. Convergent and discriminant validity of quality of life measures used in burn populations. *Burns*, 43(1):84–92.

Raina M Merchant, David A Asch, Patrick Crutchley, Lyle H Ungar, Sharath C Guntuku, Johannes C Eichstaedt, Shawndra Hill, Kevin Padrez, Robert J Smith, and H Andrew Schwartz. 2019. Evaluating the predictability of medical conditions from social media posts. *PLoS One*, 14(6):e0215476.

Steven Mesquiti, Danielle Cosme, Erik Nook, Emily Falk, and Shannon Burns. 2026. Language-based assessments can predict psychological and subjective well-being. *Communications Psychology*, 4.

G J Meyer, S E Finn, L D Eyde, G G Kay, K L Moreland, R R Dies, E J Eisman, T W Kubiszyn, and G M Reed. 2001. Psychological testing and psychological assessment. a review of evidence and issues. *Am Psychol*, 56(2):128–165.

August Nilsson, Ryan Boyd, Adithya V Ganesan, Oscar Kjell, Syeda Mahwish, Haitao Huang, Richard Rosenthal, Lyle Ungar, and H. Schwartz. 2024.

Language-based assessments for experienced well-being: Accuracy and external validity across behaviors, traits, and states.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Chong Peng and Qiang Cheng. 2021. Discriminative ridge machine: A classifier for high-dimensional data or imbalanced data. *IEEE Trans Neural Netw Learn Syst*, 32(6):2595–2609.

Duc-Anh Phan, Hiroyuki Shindo, and Yuji Matsumoto. 2016. Multiple emotions detection in conversation transcripts. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 85–94, Seoul, South Korea.

Martin Prince, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R Phillips, and Atif Rahman. 2007. No health without mental health. *Lancet*, 370(9590):859–877.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Sanne P A Rasing, Daan H M Creemers, Jan M A M Janssens, and Ron H J Scholte. 2017. Depression and anxiety prevention based on cognitive behavioral therapy for at-risk adolescents: A Meta-Analytic review. *Front Psychol*, 8:1066.

Darrel A Regier, William E Narrow, Diana E Clarke, Helena C Kraemer, S Janet Kuramoto, Emily A Kuhl, and David J Kupfer. 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry*, 170(1):59–70.

Camilo J Ruggero, Roman Kotov, Christopher J Hopwood, Michael First, Lee Anna Clark, Andrew E Skodol, Stephanie N Mullins-Sweatt, Christopher J Patrick, Bo Bach, David C Cicero, Anna Docherty, Leonard J Simms, R Michael Bagby, Robert F Krueger, Jennifer L Callahan, Michael Chmielewski, Christopher C Conway, Barbara De Clercq, Allison Dornbach-Bender, and 14 others. 2019. Integrating the hierarchical taxonomy of psychopathology (HiTOP) into clinical practice. *J Consult Clin Psychol*, 87(12):1069–1084.

Tomasz Rutowski, Elizabeth Shriberg, Amir Harati, Yang Lu, Piotr Chlebek, and Ricardo Oliveira. 2020. Depression and anxiety prediction using deep language models and transfer learning. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*, pages 1–6.

H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. 2017. DLATK: Differential language analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 55–60, Copenhagen, Denmark. Association for Computational Linguistics.

K M Scott, M Von Korff, J Alonso, M C Angermeyer, E Bromet, J Fayyad, G de Girolamo, K Demyttenaere, I Gasquet, O Gureje, J M Haro, Y He, R C Kessler, D Levinson, M E Medina Mora, M Oakley Browne, J Ormel, J Posada-Villa, M Watanabe, and D Williams. 2008. Mental-physical co-morbidity and its relationship with disability: results from the world mental health surveys. *Psychol Med*, 39(1):33–43.

Sverker Sikström, Oscar Kjell, Katarina Kjell, and H. Andrew Schwartz. 2024a. Question-based computational language assessment of psychological states. *Communications Psychology*, 2:97.

Sverker Sikström, Miriam Nicolai, Josephine Ahrendt, Suvi Nevanlinna, and Lotta Stille. 2024b. Language or rating scales based classifications of emotions: computational analysis of language and alexithymia. *npj Mental Health Research*, 3(1):37.

Troels Steenstrup, Ole Birger Pedersen, Jacob Hjelmborg, Axel Skytthe, and Kirsten Ohm Kyvik. 2013. Heritability of health-related quality of life: SF-12 summary scores in a population-based nationwide twin cohort. *Twin Res Hum Genet*, 16(3):670–678.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

8 Appendix

WTC interviews. Participants who consented to take part in research responded to questions automatically displayed on a screen in a private room during their clinical visit. Questions probed both positive and negative aspects of participants' past, present, and future lives, as well as experiences related to specific events, including 9/11 and COVID-19, using broad language rather than clinical symptom terminology. Participants were instructed not to read the questions aloud and to spend at least 60 seconds per question; recordings meeting a minimum threshold of 150 words were retained, with a mean response time of 7.5 minutes (SD = 4.1). Interviews were transcribed using Whisper-large-v2 (Radford et al., 2022) and diarized to isolate participant responses.

iHiTOP interviews. Participants engaged in a semi-structured interview with a research assistant. The interviewers asked probing questions pertaining to each overarching HiTOP construct, and assigned ratings according to participant responses. Interviewers followed a consistent rating guideline to ensure reliable outcomes for analysis.

Experimental setup. Language embeddings were extracted and analyzed from Whisper-large-v3 transcripts (Radford et al., 2022) using Layer 23 from RoBERTa-large (Liu et al., 2019) via the Python library Differential Language Analysis Toolkit (DLATK; Schwartz et al. (2017)). Embedding dimensions were standardized to zero mean and unit variance prior to model fitting. Ridge Regression models were then trained separately for each outcome using ten-fold cross-validation, with the discriminant penalty strength γ varied across a predefined grid to characterize the trade-off between convergent and discriminant validity.