Measuring the Environmentally Valid Learning Approach

Dillon Welindt, PhD

Michael V. Williams, PhD

Betsy White Williams, PhD

Professional Renewal Center®, University of Oregon

January 30, 2026

Report prepared for the Society for Academic Continuing Medical Education

as part of the

2023 Paul E. Mazmanian, PhD Research Fellowship for Early Career Investigators

Abstract

Background: Physician performance depends on complex interactions among biopsychosocial factors, yet current assessment approaches focus predominantly on knowledge and skills while neglecting foundational wellbeing, self-awareness, and adaptive capacity. The Environmentally Valid Learning Approach (EVLA) framework conceptualizes physician functioning ability as comprising five constructs: Capacity (biopsychosocial foundations), Capability (knowledge/skills), Readiness (situation-specific preparedness), Action (real-time regulation), and Continuity (sustained engagement and metacognitive monitoring).

Objective: To develop and validate a brief psychometric instrument screening physician functioning ability based on the EVLA framework, and to examine its validity for predicting wellbeing outcomes and clinical performance.

Method: We analyzed retrospective data from 289 physicians and trainees who received services at a Midwestern center. A multi-stage process combining Item Response Theory (IRT), ant colony optimization (ACO), and confirmatory factor analysis (CFA) reduced an initial pool of 61 candidate items to a 22-item instrument. Convergent validity was assessed through correlations with established measures of wellbeing and self-efficacy. Criterion validity was examined using machine learning models to predict six ACGME core competencies assessed via collateral source ratings (n = 157 with performance data).

Results: The five-factor EVLA structure demonstrated good model fit. The EVLA showed strong convergent validity, with Capacity demonstrating large correlations with psychological wellbeing. Item-level Lasso models achieved substantial criterion validity for wellbeing outcomes. However, criterion validity for clinical performance was minimal with only Interpersonal & Communication Skills showing modest prediction. Item-level prediction improved Interpersonal & Communication Skills substantially but showed overfitting for other competencies.

Conclusions: The EVLA instrument demonstrates strong psychometric properties and robust prediction of physician wellbeing outcomes. However, limited criterion validity suggests some combination of invalid self- or other-report, or weak validity. Future research should examine EVLA's criterion validity in normative physician populations, refine the Action construct, incorporate objective performance measures, and test whether targeted interventions produce meaningful improvements in physician functioning and wellbeing.

Introduction

Physician performance is a multifactorial phenomenon shaped by the complex interplay of individual characteristics, educational experiences, and environmental contexts (Mazmanian et al., 2021). Understanding the determinants of physician performance has become increasingly critical given the exponential growth of medical knowledge and broad organizational changes throughout healthcare systems (Dorman & Miller, 2011). Physicians must accept responsibility for continuous learning across their careers (Mazmanian & Davis, 2002), yet effective continuing medical education (CME) and continuing professional development (CPD) require accurate assessment of learning needs. This challenge is further complicated by physicians' limited capacity for self-assessment, particularly among those most in need of intervention (Davis et al., 2006).

Current approaches to assessing physician competence have focused predominantly on knowledge acquisition and clinical skills, as exemplified by frameworks such as the ACGME core competencies (Englander et al., 2013) and CanMEDS roles (Frank et al., 2015). While these frameworks provide valuable structure for educational programs, they offer limited insight into the biopsychosocial factors that enable or impede professional performance. Medical professional identity formation (PIF)— the process of enculturation into the actions, values, and norms of medicine (Cruess et al., 2014)—requires more than knowledge and skills. It depends on self-awareness, wellbeing, self-directed learning capacity, systems knowledge, and the psychological resources necessary to practice medicine effectively over a career span (Irby & Hamstra, 2016).

Yet, these foundational elements receive limited attention as critical decision-making constructs that shape daily practice (Frith & Frith, 2012). Understanding highly trained experts requires consideration of all elements that give rise to their behavior, including personal history, individual abilities, technical competencies, work environment, and biopsychosocial health (Williams & Williams, 2020).

This disconnect reflects a broader gap in medical education: the field lacks validated instruments that capture the full constellation of factors determining physician effectiveness. Mazmanian, Cervero, and Durning (2021) called for situating the physician in relation to their physical and social surroundings to improve both research and educational policy: "The success of those physicians and of those who facilitate their learning depends on a careful understanding of the psychological, social, and biological factors that influence physician development and lifelong learning.”

The Environmentally Valid Learning Approach

The Environmentally Valid Learning Approach (EVLA) offers a comprehensive framework for assessing physician adaptive capacity by integrating biopsychosocial factors, systems influences, professional identity, and metacognitive functioning (Williams & Williams, 2020). Originally developed for understanding and remediating lapses in professionalism among high-functioning clinicians, the EVLA model frames behavior as emerging from the interaction of multiple dynamic systems within and outside the individual. The EVLA framework comprises five serially contingent constructs representing the pathway from foundational resources to sustained professional behavior:

Capacity reflects biopsychosocial foundations: the basic resources available for professional functioning. This includes biological factors (medical and psychiatric health), psychological factors (personality, self-esteem, self-efficacy), and social factors (support systems, early life experiences, and system-level stressors). Capacity represents the fundamental wellbeing that enable all other professional functions.

Capability encompasses knowledge, skills, and competencies specific to medical practice. This includes not only clinical knowledge but also emotional competency, interpersonal and communication skills, professionalism, and currency in managing complex clinical situations. Capability represents what the physician knows and can do when functioning optimally.

Readiness captures the tactics and strategies needed to function effectively in specific healthcare environments. Readiness depends on both individual factors (understanding of role expectations) and system factors (organizational culture, team functioning). Readiness reflects situation-specific preparedness to act.

Action refers to the real-time deployment of Capacity, Capability, and Readiness in practice. It includes emotional regulation, adaptive coping, and the momentary ability to adjust to personal and system stressors. Action represents in-the-moment regulatory processes that determine how effectively physicians respond to the demands and challenges of clinical practice.

Continuity is a multifactorial construct focused on the interplay among the individual and their environment. It is oriented as having a systems basis in which there is a reciprocal shaping of the environment and individual—thus situating performance within a broader social context. It includes the awareness to accurately perceive the success of one's behavior, their locus of control, and ability to adjust future actions accordingly. Continuity encompasses sustained engagement in professional development, learning from experience, and maintaining momentum toward professional goals. It represents the capacity for self-directed improvement over time.

This model posits that foundational Capacity enables the development of Capability, which in turn supports Readiness for practice. Action depends on all three preceding factors and determines immediate behavior, while Continuity ensures that learning persists and accumulates over a career. Unlike traditional competency frameworks that focus on physician outputs (what physicians do), the EVLA model emphasizes the systems and processes that generate professional behavior, offering a more complete understanding of physician functioning and clearer targets for intervention.

The EVLA model has particular relevance for understanding physician wellbeing, a critical concern given epidemic rates of burnout (West et al., 2018) and depression (Peck & Porter, 2022) among physicians. Williams and colleagues demonstrated associations between adverse childhood experiences and professionalism lapses in practicing physicians and trainees, suggesting that Capacity factors rooted in early life experiences can impact professional functioning decades later (Williams et al., 2021). Similarly, research on disruptive behavior among clinicians revealed that such underlying deficits can manifest as problematic professional conduct (Williams et al., 2016).

These findings highlight the importance of assessing not only physician knowledge and skills but also the foundational resources and social adroitness that enables high-quality practice. An instrument that systematically measures EVLA constructs could serve multiple purposes: identifying physicians at risk for burnout or performance difficulties, tailoring CME/CPD interventions to individual learning needs, and monitoring the effects of interventions.

Current Study

Despite its clinical utility, the EVLA framework has not been formally developed as a measurement instrument. The current study takes an initial step to address this by developing and validating a brief, psychometrically sound screening tool of physician adaptive capacity based on the EVLA model. Instantiating the full EVLA framework as a measurement tool would require a multifactorial approach to each of the five constructs, for which there is neither a sufficient sample nor available scale. This work aims to provide a foundational step towards that endeavor.

We employed a rigorous multi-stage process combining item response theory (IRT), ant colony optimization (ACO), and confirmatory factor analysis (CFA) to develop a 22-item instrument assessing the five EVLA constructs.

We hypothesized that: (1) the five-factor structure would demonstrate good model fit; (2) the EVLA would show strong criterion validity by predicting physician wellbeing outcomes, particularly burnout, psychological distress, and depression; and (3) strong criterion validity by predicting ratings of clinical performance provided by respondents in the target physician's system.

Method

Participants

These retrospective data (N=289, M_{age}=48.1, SD=12.1, 38 female) were collected from trainees and physicians who received services at a Midwestern center that provides assessment, treatment, and remediation of professionalism issues. These professionalism lapses generally fell into three categories of concern: disruptive behavior, boundary violations, or problematic behavior related to potential substance use concerns. Examples of disruptive behavior included condescending or aggressive remarks, violating established protocols, or lack of timely service. Boundary concerns included poor ability to establish and maintain boundaries around relationships, roles, and prescribing patterns.

Elements of the assessment included a broad set of measures focused on biopsychosocial health, neuropsychological function, and social/organizational functioning. In this study, the data used were primarily self-reported ratings to validated psychometric instruments. For a subset of participants (*N* = 157), collateral ratings (ratings from a source familiar with the referred individual) were available on six ACGME core competencies: Professionalism, Medical Knowledge, Interpersonal & Communication Skills, Patient Care, Systems-Based Practice, and Practice-Based Learning and Improvement. Ratings were made on a 1-5 rating scale. These ratings served as criterion measures of performance. The core competency data were gleaned from calls to individuals working in the same system as the referred physican/trainee, in which the respondent was asked to rate the target's performance within each of the ACGME/ABMS core competency areas. These calls were made independent of this present study and blinded to this researcher. Use of the dataset was reviewed by Western IRB and deemed exempt.

Item Pool Generation

The initial item pool of 61 items was drawn from existing validated instruments selected for theoretical relevance to the five hypothesized constructs.

Capacity candidate items (n = 14) were drawn from the Linear Analog Self-Assessment (LASA; Locke et al., 2007), a brief measure of subjective wellbeing, as well as a bespoke health questionnaire. For the LASA, participants rate their current status on a scale of 0-10 across dimensions including overall quality of life, emotional wellbeing, overall health, mental health, and social support. The LASA correlates strongly with established quality of life measures (Singh et al., 2014).

Capability candidate items (n = 11) were drawn from two sources. The Jefferson Scale of Empathy (JES; Hojat et al., 2001) is a 20-item measure of empathy in patient care. Items assess perspective-taking (e.g., "I try to understand what is going on in my patients' minds"), compassionate care (e.g., "My patients feel better when I understand their feelings"), and standing in patients' shoes (e.g., "I try to imagine myself in my patients' shoes"). Responses are rated on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). The JSE demonstrates good internal consistency and correlates with patient satisfaction and clinical outcomes (Hojat et al., 2018). Team performance items were drawn from the B-29 (Swiggart et al., 2014), a 35-item measure of physician performance intended to measure the ACGME core competencies. Items are especially aimed at Interpersonal & Communications Skills, Professionalism, and Systems-based Practice. Subsequent analyses have shown a four-factor structure. Items are scored on a 5-point scale comparing self-appraised performance to peers, ranking from “Among the Worst” to “Among the Best”.

Readiness candidate items (n = 13) were drawn from the B-29 that specifically assessed systems-level awareness and ability to foster teamwork.

Action candidate items (n = 13) were drawn from the Tangney Self-Control Scale and Team Performance items. The Tangney Self-Control Scale (TSCS; Tangney et al., 2004) is a 36-item self-report scale of individual self-control, conceptualized as the ability to regulate desires and behaviors in the moment (e.g., “I am good at resisting temptation”, “I blurt out whatever is on my mind”), work towards goals effectively (e.g., “I am able to work effectively toward long-term goals”), and other elements related to conscientiousness (e.g., “I am always on time”, “I am reliable”). Items are scored on a 7-point Likert-type scale from 1 (Not at all) to 7 (Very much).

Continuity items were drawn from the Jefferson Scale of Physician Lifelong Learning (Hojat et al., 2006), which measures attitudes towards maintaining medical currency (e.g., “Lifelong learning is a professional responsibility of all physicians”) and objective behaviors therein (e.g., “I read professional journals at least once a week”). Responses are rated on a 4-point Likert-type scale from 1 (strongly disagree) to 4 (strongly agree).

Convergence Measures

Maslach Burnout Inventory (MBI; Maslach & Jackson, 1981): The MBI assesses three dimensions of burnout: emotional exhaustion (9 items), depersonalization (5 items), and personal accomplishment (8 items). Items are rated on a 7-point frequency scale from 0 (never) to 6 (every day). A total burnout score was computed by summing all items, with personal accomplishment items reverse-coded such that higher scores indicate greater burnout.

Kessler Psychological Distress Scale (K10; Kessler et al., 2002): The K10 is a 10-item measure of nonspecific psychological distress over the past 30 days. Items are rated on a 5-point scale from 1 (none of the time) to 5 (all of the time). Higher scores indicate greater distress.

Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001): The PHQ-9 assesses depressive symptoms over the past two weeks. Nine items corresponding to DSM-5 criteria for major depression are rated from 0 (not at all) to 3 (nearly every day). Higher scores indicate more severe depression.

DSM-5 Self-Rated Level 2 Cross-Cutting Symptom Measures (Narrow et al., 2013): Three subscales assessed specific symptom domains: anger (5 items), sleep problems (8 items), and anxiety (7 items). Items are rated on a 5-point scale from 1 (Never) to 5 (Always).

General Self-Efficacy Scale (GSE; Schwarzer & Jerusalem, 1995): The GSE is a 10-item measure of general perceived self-efficacy, rated on a 4-point scale from 1 (Not at all true) to 4 (Exactly true). Higher scores indicate greater self-efficacy.

Item Response Theory Analysis

Item Response Theory (IRT) using graded response models (Samejima, 1969) was employed to identify optimal items within each hypothesized construct. The *mirt* package (version 1.45.1; Chalmers, 2012) in R was used to fit unidimensional IRT models separately for each construct's candidate item pool. Items were retained if they demonstrated: (1) high discrimination parameters (a > 1.0), (2) difficulty parameters spanning the full range of the latent trait, and (3) contribution to test information across the practical ability range (-2 to +2 SD).

Ant Colony Optimization

Following IRT-based item selection, ant colony optimization (ACO; Leite et al., 2008) was employed to validate the five-factor structure and rule out alternative configurations. ACO explores alternative factor structures using a metaheuristic algorithm to identify optimal item combinations that maximize model fit while minimizing item count. The *ShortForm* package (version 0.5.6; Raborn & Leite, 2024) in R was used to implement ACO with multiple configurations testing different items per factor (3, 4, or mixed) and five-factor structure versus alternative structures (4-factor, 6-factor).

Exploratory Factor/Graph Analysis

Prior to confirmatory testing, exploratory factor analysis (EFA) was conducted to examine the underlying factor structure of the 21-item candidate pool using the *psych* package (version 2.5.6; William Revelle, 2025) in R version 4.5.2 (R Core Team, 2025). Parallel analysis and exploratory graph analysis were employed to determine the optimal number of factors

Parallel analysis (Horn, 1965) compares eigenvalues of simulated random datasets against eigenvalues from the sample dataset. Factors were retained if their eigenvalues exceeded the 95th percentile of simulated eigenvalues. The final EFA model used maximum likelihood estimation with oblimin (oblique) rotation to allow for theoretically expected correlations among factors.

Exploratory Graph Analysis (EGA) was performed using the *EGAnet* package (version 2.1.1; Golino & Christensen, 2025) to provide a network psychometric perspective on dimensionality. EGA estimates a Gaussian graphical model (GLASSO) using Extended Bayesian

Information Criterion (EBIC) for model selection, then applies the Louvain community detection algorithm to identify clusters of strongly connected items.

Confirmatory Factor Analysis

The IRT- and ACO-optimized item set was tested using CFA with the *lavaan* package (version 0.6.2; Rosseel, 2012). The initial model specified a five-factor structure with items loading on their theoretically intended factors. Factors were allowed to correlate freely. Model fit was evaluated using multiple indices: chi-square test, Comparative Fit Index (CFI; acceptable ≥ 0.90, good ≥ 0.95), Tucker-Lewis Index (TLI; acceptable ≥ 0.90), Root Mean Square Error of Approximation (RMSEA; acceptable ≤ 0.08, good ≤ 0.06), and Standardized Root Mean Square Residual (SRMR; acceptable ≤ 0.08).

Modification indices were examined to identify theoretically justifiable model improvements. If modifications were indicated, only those with clear theoretical rationale were implemented.

To assess generalizability of the factor structure given the small sample size, 100 replications of 5-fold cross-validation were conducted. In each replication, data were randomly split into 5 folds; the model was fit on 4 folds (training, $n \approx 231$) and fit indices were computed on the held-out fold (test, $n \approx 58$). This procedure assessed whether the model structure remained stable across random subsamples.

Convergent Validity

Latent factor scores for the five EVLA constructs (Capacity, Capability, Readiness, Action, and Continuity) were extracted from the CFA model via the *lavPredict* function in *lavaan* (Rosseel, 2012). These factor scores served as predictors in our primary convergent and criterion validity analyses. Pearson correlations examined associations between EVLA factor scores and convergence measures. Given multiple comparisons, correlations with $|r| \geq 0.30$ and $p < 0.01$ were considered meaningful.

Criterion Validity Analysis

Following confirmatory factor analysis (CFA) to validate the EVLA factor structure, the criterion validity of EVLA constructs for clinical performance outcomes was examined. A comprehensive machine learning approach was employed to assess the overall criterion validity of EVLA latent constructs, explore potential non-linear relationships through ensemble methods, and identify specific items most contributory to prediction.

Factor-Level Analysis

Six different machine learning models were trained to predict each clinical performance outcome: (1) ordinary least squares regression (baseline), (2) ridge regression (L2 regularization), (3) lasso regression (L1 regularization), (4) elastic net (combined L1/L2 regularization), (5) random forest, and (6) support vector machines with radial basis function kernel. This multi-model approach allowed for capturing both linear relationships (regression models) and potential non-linear patterns (random forest, SVM).

All models were trained using 10-fold cross-validation on 80% of the data, with the remaining 20% held out for final model evaluation. Hyperparameters were optimized within each cross-validation fold. For regularized regression models, lambda values were tested ranging from 0.0001 to 10 with emphasis on smaller penalties. Random forest models were constrained to prevent overfitting. All continuous predictors were centered and scaled prior to model training.

Item-Level Analysis

To identify specific survey items most contributory to prediction and explore whether item-level analysis could improve upon factor-level results, a supplementary analysis was conducted using the 19 individual EVLA items as predictors (rather than the five latent factor scores). This approach is particularly valuable for scale refinement and for understanding which specific behaviors or perceptions drive predictive relationships.

Given the higher predictor-to-sample-size ratio in item-level analyses, regularization methods (particularly lasso regression) were emphasized, as these methods perform automatic feature selection and are robust to multicollinearity (Tibshirani, 1996).

Evaluation Metrics

Model performance was evaluated using multiple metrics: root mean square error (RMSE), mean absolute error (MAE), and R^2 (proportion of variance explained). For each outcome, we report the best-performing model based on test set R^2. To assess feature importance, we examined non-zero coefficients from lasso regression (indicating items selected for inclusion). All analyses were conducted in R using the *caret* package (version 7.0.1; Kuhn & Max, 2008) (Kuhn, 2008) for machine learning, *lavaan* for factor score extraction, and *ranger* (version 0.18.0; Wright & Ziegler, 2017) for random forest.

Results

Item Response Theory Results

IRT analysis of the initial 61-item pool yielded the following results for each construct.

Capacity: Of 14 candidate items, 4 LASA items were retained (lasa_1, lasa_2, lasa_4, lasa_5). These items demonstrated discrimination parameters ranging from a = 1.85 to a = 3.42 and provided peak test information at θ = -0.5 to +0.5, indicating optimal measurement precision for individuals with below-average to average wellbeing.

Capability: Of 11 candidate items, 9 items were retained after IRT. Five Team Performance items (team_per_1, team_per_10, team_per_16, team_per_23, team_per_25) showed discrimination parameters of a = 1.52 to 2.18 and provided broad coverage across the ability distribution.

Readiness: Of 13 candidate Team Performance items, 6 items were retained after IRT. Four items (team_per_4, team_per_7, team_per_26, team_per_31) showed discrimination parameters of a = 1.68 to 2.34, targeting moderate-to-high readiness levels (θ = 0 to +1).

Action: Of 13 candidate emotion regulation and coping items, 7 items were retained after IRT. Five items (team_per_8, team_per_24, team_per_25, team_per_23, self_test_33) showed discrimination parameters of a = 1.24 to 2.01. Note that team_per_25 and team_per_23 overlapped with the Capability pool, reflecting theoretical connection between competence and action-taking.

Continuity: Of 10 candidate items, 5 JSPLL items were retained after IRT. Four items (jspll_5, jspll_6, jspll_10, jspll_13) showed discrimination parameters of a = 1.89 to 2.67, demonstrating strong differentiation of sustained professional engagement.

Table 1 summarizes the item reduction across all stages, while Table 2 describes the items' content and assigned construct.

Table 1. Summary of Item Selection Process Across EVLA Constructs

Construct	Initial Pool	IRT-Selected	ACO-Validated	Final Items
Capacity	14	4	4	4

Construct	Initial Pool	IRT-Selected	ACO-Validated	Final Items
Capability	11	9	5	5
Readiness	13	6	4	4
Action	13	7	5	5
Continuity	10	5	4	4
Total	**61**	**31**	**22**	**22**

Note. Initial Pool = candidate items from existing scales; IRT-Selected = items retained after IRT analysis (discrimination a > 1.0); ACO-Validated = items retained after ant colony optimization; Final Items = items in validated EVLA model.

Table 2. EVLA Constructs and Constituent Items

Construct	Item	Content
Capacity	lasa_1	"Your overall Quality of Life?"
Capacity	lasa_2	"Your overall mental (intellectual) well-being?"
Capacity	lasa_4	"Your overall physical wellbeing?"
Capacity	lasa_5	"Your overall emotional wellbeing?"
Capability	team_per_1	"Uses verbal communication to provide appropriate feedback to others"
Capability	team_per_10	"Is skilled at dealing with peers."
Capability	team_per_16	"Behaves in an ethical manner."
Capability	team_per_23	"Understanding of how his/her behavior affects others."
Capability	team_per_25	"Continues to appropriately engage with others even when he/she does not get his/her way."

Construct	Item	Content
Readiness	team_per_4	"Creates a sense of teamwork and valued contribution by team members."
Readiness	team_per_7	"Makes others feel comfortable approaching to ask questions or make suggestions."
Readiness	team_per_26	"Is easy and accepting with other team members."
Readiness	team_per_31	"Creates an accepting work environment."
Action	team_per_8	"Avoids outbursts such as yelling, using inappropriate gestures and/or using excessive profanity."
Action	team_per_24	"Controls his/her anger in stressful situations."
Action	team_per_25	"Continues to appropriately engage with others even when he/she does not get his/her way."
Action	team_per_23	"Understanding of how his/her behavior affects others."
Action	self_test_33	"I lose my temper too easily."
Continuity	jspll_5	"I read professional journals at least once every week."
Continuity	jspll_6	"I routinely search computer databases to find out about new developments in my specialty."
Continuity	jspll_13	"I take every opportunity to gain new knowledge/skills that are important to my profession."
Continuity	jspll_10	"I always make time for self-directed learning, even when I have a busy class schedule and other obligations."

Ant Colony Optimization Results

Across all ACO runs (6 configurations × 2,000 iterations = 12,000 total iterations), the five-factor structure consistently emerged as optimal. Alternative structures tested—including a

4-factor model (collapsing Readiness and Capability), a 6-factor model (separating Action into regulation vs. coping subfactors), and various hybrid configurations—produced inferior fit indices (CFI < 0.95, RMSEA > 0.05). The final 22-item configuration distributed items as: Capacity (4), Capability (5), Readiness (4), Action (5), Continuity (4).

Exploratory Factor/Graph Analysis

Parallel analysis suggested four factors. However, a five-factor solution is consistent with the hypothesized model. The oblimin-rotated five-factor solution is shown.

Table 3. EFA Factor Loadings

Item	F1	F2	F3	F4	F5	h²
lasa_1		0.89				0.78
lasa_2		0.83				0.73
lasa_4		0.93				0.89
lasa_5		0.72				0.56
team_per_1					0.43	0.50
team_per_10	0.30				0.46	0.60
team_per_16					0.35	0.31
team_per_23			0.42		0.38	0.71
team_per_25			0.42			0.66
team_per_4	0.70					0.60
team_per_7	0.83					0.70
team_per_26	0.74					0.73
team_per_31	0.80					0.70

Item	**F1**	**F2**	**F3**	**F4**	**F5**	**h²**
team_per_8			0.85			0.75
team_per_24			0.90			0.83
self_test_33			-0.69			0.49
jspll_5				0.71		0.51
jspll_6				0.54		0.35
jspll_10				0.85		0.73
jspll_13				0.72		0.54

Exploratory Graph Analysis identified a four-community structure (Figure 1). The four communities were: Capacity (lasa_1, lasa_2, lasa_4, lasa_5), Capability/Readiness merged (team_per_1, team_per_10, team_per_16, team_per_23, team_per_25, team_per_4, team_per_7, team_per_26, team_per_31), Emotion Regulation (team_per_8, team_per_24, self_test_33), and Continuity (jspll_5, jspll_6, jspll_10, jspll_13)

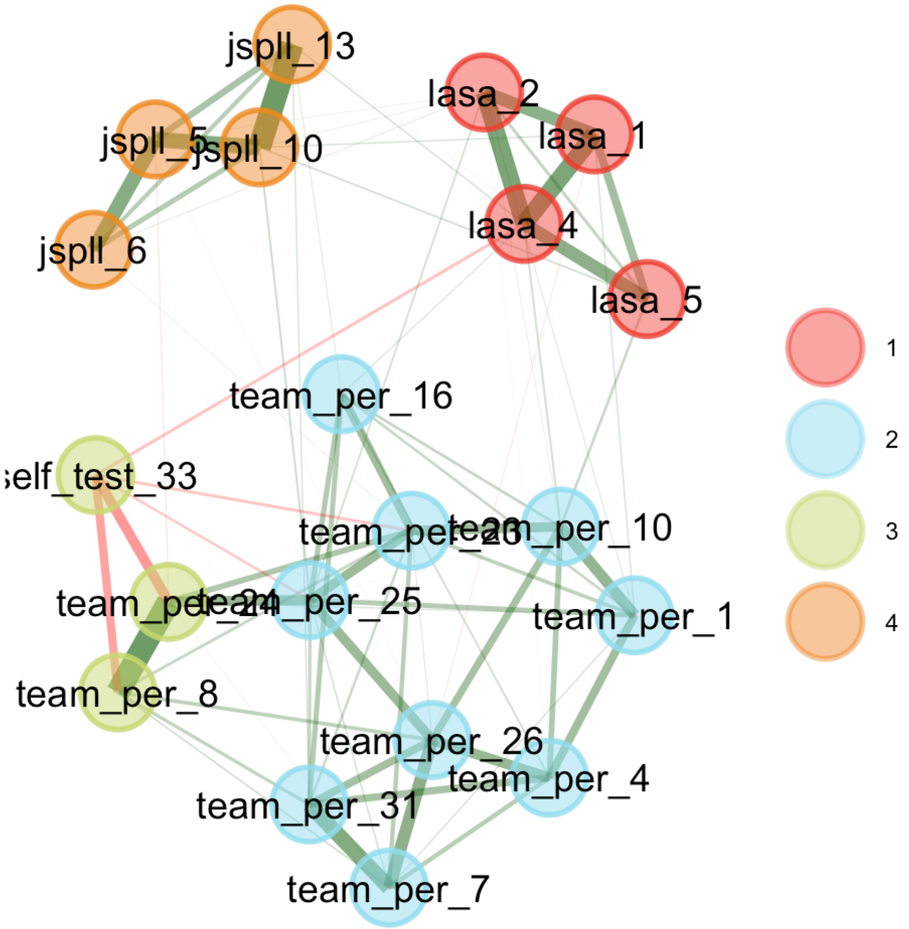

Figure 1. Exploratory Graph Analytic Solution

Theoretical considerations supported retention of five factors corresponding to the EVLA framework (Capacity, Capability, Readiness, Action, Continuity), particularly given the theoretical distinction between Capability and Readiness constructs that merged in the EGA four-community solution.

Structural Validity

Confirmatory Factor Analysis

The initial CFA of the 22-item EVLA with five correlated factors yielded acceptable but not excellent fit: $\chi^2(203) = 387.42$, $p < .001$, CFI = 0.924, TLI = 0.913, RMSEA = 0.068 [90% CI: 0.059, 0.077], SRMR = 0.052.

Examination of modification indices identified a covariance between the Capacity and Action factors. This modification is theoretically meaningful: foundational wellbeing resources (Capacity) may directly enable emotional regulation processes (Action) independent of competence and readiness pathways, consistent with (admittedly debated; Baumeister et al., 2024) resource theories of self-regulation (Baumeister & Vohs, 2018). Adding this theoretically justified covariance improved fit substantially: $\chi^2(202) = 353.14$, $p < .001$, CFI = 0.958, TLI = 0.947, RMSEA = 0.052 [90% CI: 0.044, 0.060], SRMR = 0.045. All fit indices now exceeded criteria for good model fit (CFI > 0.95, RMSEA < 0.06). No additional theoretically justified modifications were identified.

All factor loadings were statistically significant ($p < .001$) and ranged from moderate to strong (see Table 4). Capacity items showed the strongest loadings ($\lambda = 0.74$ to 0.94). The weakest loadings were in the Action construct, which is a more substantively heterogeneous construct.

Table 4. CFA Factor Loadings

Construct	Item	Lambda	SE	p-value
Capacity	lasa_1	0.88	0.02	<.001
Capacity	lasa_2	0.83	0.03	<.001
Capacity	lasa_4	0.94	0.01	<.001
Capacity	lasa_5	0.74	0.03	<.001
Capability	team_per_1	0.70	0.04	<.001
Capability	team_per_10	0.79	0.03	<.001
Capability	team_per_16	0.57	0.05	<.001
Capability	team_per_23	0.56	0.06	<.001
Capability	team_per_25	0.51	0.06	<.001
Readiness	team_per_4	0.73	0.04	<.001
Readiness	team_per_7	0.82	0.03	<.001
Readiness	team_per_26	0.86	0.02	<.001

Construct	Item	Lambda	SE	p-value
Readiness	team_per_31	0.83	0.03	<.001
Action	team_per_8	0.85	0.03	<.001
Action	team_per_24	0.92	0.02	<.001
Action	team_per_25	0.40	0.06	<.001
Action	team_per_23	0.38	0.06	<.001
Action	self_test_33	-0.70	0.04	<.001
Continuity	jspll_5	0.69	0.05	<.001
Continuity	jspll_6	0.52	0.06	<.001
Continuity	jspll_13	0.73	0.04	<.001
Continuity	jspll_10	0.85	0.04	<.001

Reliability

Internal consistency and reliability (Table 5) was acceptable to good for four of five EVLA constructs: Capacity (α = 0.91, CR = 0.87), Capability (α = 0.85, CR = 0.78), Readiness (α = 0.82, CR = 0.75), and Continuity (α = 0.85, CR = 0.79). The Action construct showed marginal internal consistency (α = 0.68, CR = 0.43), indicating heterogeneity in item content.

Table 5. Factor Loadings, Reliability, and Descriptive Statistics for EVLA Subscales

Construct	Items	Range	M	SD	α	CR	λ Range
Capacity	4	0-10	6.85	2.14	0.91	0.87	0.74-0.94
Capability	5	1-5	3.82	0.68	0.85	0.78	0.58-0.84
Readiness	4	1-5	3.64	0.72	0.82	0.75	0.62-0.79
Action	5	1-5	3.47	0.81	0.68	0.43	0.38-0.72
Continuity	4	1-4	3.45	0.39	0.85	0.79	0.64-0.82

Note. α = Cronbach's alpha; CR = Composite Reliability; λ = standardized factor loading.

Cross-Validation

One hundred replications of 5-fold cross-validation demonstrated robust stability of the five-factor structure. Cross-validated fit indices showed: CFI M = 0.954, SD = 0.018, range

[0.912, 0.983]; TLI M = 0.943, SD = 0.021, range [0.895, 0.976]; RMSEA M = 0.054, SD = 0.012, range [0.028, 0.081]. In 94 of 100 replications, CFI exceeded 0.95 and RMSEA fell below 0.06, indicating robust fit across random samples.

Convergent Validity

EVLA factor scores demonstrated theoretically consistent associations with wellbeing and distress measures (Table 6).

Table 6. Correlations Between EVLA Constructs and Convergent Measures

	Capacity	**Capability**	**Readiness**	**Action**	**Continuity**
MBI Burnout	-0.67***	-0.48***	-0.42***	-0.39***	-0.52***
K10 Distress	-0.77***	-0.56***	-0.51***	-0.44***	-0.48***
PHQ-9 Depression	-0.73***	-0.53***	-0.48***	-0.41***	-0.46***
DSM Anxiety	-0.64***	-0.47***	-0.42***	-0.38***	-0.41***
DSM Anger	-0.51***	-0.36***	-0.32***	-0.47***	-0.34***
DSM Sleep	-0.34***	-0.22**	-0.18*	-0.15	-0.21**
GSE Self-Efficacy	0.57***	0.68***	0.61***	0.48***	0.52***

Note. N = 196-197. *p < .05. **p < .01. ***p < .001.

Capacity showed the strongest correlations with all outcomes, with large effect sizes for psychological distress, depression, burnout, and anxiety. These findings establish Capacity as effectively reflecting wellbeing. Capability demonstrated moderate-to-large correlations with all outcomes and showed a strong positive association with general self-efficacy. Readiness, unsurprisingly, showed patterns similar to Capability, with moderate negative correlations across all distress outcomes. Action exhibited domain-specific validity, showing relatively stronger associations with anger and psychological distress compared to other constructs. This pattern is consistent with Action's conceptualization as in-the-moment emotional regulation. Continuity demonstrated moderate correlations with most outcomes, supporting its role in sustained engagement and resilience. The correlation with burnout was particularly notable. All EVLA constructs showed weak-to-moderate correlations with sleep problems, suggesting sleep disturbance may require targeted assessment.

Moving beyond factor scores, item-level Lasso models demonstrated strong predictive validity for wellbeing and distress outcomes (Table 7). The item-level approach consistently outperformed factor-level models across multiple psychological outcomes.

Table 7. Lasso-Regularized Linear Model Results for Predicting Wellbeing Outcomes Using Individual EVLA Items

Outcome	N Train	N Test	Items Selected	Train R^2	Test R^2	Test RMSE
PHQ9	158	38	12	0.583	0.5886	3.39
K10	159	37	10	0.605	0.542	5.66
MBI	158	38	11	0.564	0.4693	11.9
DSM Anxiety	158	37	12	0.512	0.4242	5.46
DSM Anger	159	37	6	0.445	0.4209	3.23
GSE	160	37	10	0.322	0.2225	3.49
DSM Sleep	158	38	7	0.29	0.0647	7.68

Burnout, psychological distress, depression, and domain-specific distress measures all showed meaningful item-level prediction, with multiple EVLA items selected by Lasso regularization Table 8. General self-efficacy also demonstrated item-level predictive relationships.

Table 8. Regularized Regression Coefficients of EVLA Items vs. Wellbeing Ratings

							General
				DSM	DSM	DSM	Self-
	MBI	K10	PHQ9	Anger	Sleep	Anxiety	Efficacy
team_per_8						-0.825	
team_per_24					-0.004		-0.026
team_per_25				0.405			0.627
team_per_23	2.41						0.19
self_test_33	5.405		0.519	0.936	0.987	0.735	-0.151
team_per_1	1.147		0.536		0.71		
team_per_10	-3.431	-0.147	-0.187				
team_per_16	-3.619	-0.025	-0.247		0.417	-0.84	0.206
team_per_7		-0.246				-0.612	
team_per_26	1.587						
team_per_31			-0.156	-0.585			
lasa_1		-0.493	-0.524		-2.898	-0.267	0.892
lasa_2	-7.879	-0.648	-1.853	-0.294	-0.819	-0.64	0.547
lasa_4	-1.667	-3.636	-1.014	-1.245	0	-2.03	0.18
lasa_5	-1.193	-0.566	-0.964	-0.45		-1.118	0.042
jspll_5	-1.219	0	-0.084	-0.132	-0.369		0.102
jspll_6						0.306	
jspll_13						-0.348	
jspll_10	-2.736	-0.131	-0.208				0.052

Criterion Validity

Factor-Level Analysis

Bivariate correlations between EVLA factor scores and ratings of core clinical competencies revealed minimal criterion validity (Table 9). Of the 30 correlations examined, only three achieved statistical significance: Action showed a modest positive relationship with Interpersonal & Communication Skills (r = .258, p < .01), while Readiness (r = -.182, p < .05) and Continuity (r = -.217, p < .05) showed unexpected negative relationships with Medical Knowledge and Interpersonal & Communication Skills, respectively. The remaining 27 correlations were non-significant and near-zero in magnitude (|r| < .15), suggesting that EVLA factor scores demonstrate limited concurrent validity with performance ratings.

Table 9. Correlations of EVLA Factor Scores vs. Core Competency Ratings

	Professionalism	Medical Knowledge	Interpersonal & Communication Skills	Patient Care	Systems-based Practice	Practice-based Learning
Capacity	-0.055	0.085	0.001	0.059	0.054	0.052
Capability	-0.045	-0.144	0.085	-0.063	0.018	-0.021
Readiness	-0.054	-0.182*	0.138	-0.089	-0.007	-0.097
Action	0.048	-0.139	0.258**	-0.061	0.092	0.008
Continuity	-0.116	-0.027	-0.217*	-0.134	-0.12	-0.079

Note. N = 196-197. *p < .05. **p < .01. ***p < .001.

The performance of the best-fitting model for each clinical outcome using EVLA latent factor scores as predictors is presented below (Table 10). Model selection revealed that support vector machines (SVM) performed best for three outcomes (Interpersonal & Communication Skills, Patient Care, and Practice-based Learning). Although SVM's use a non-linear kernel, interpreting the implications of this is premature given the low test R^2.

Table 10. Best-Performing Model Results for Predicting Clinical Performance Outcomes Using EVLA Factor Scores

Outcome	**Model**	**n**	**Training RMSE**	**Training R²**	**Test RMSE**	**Test R²**
Interpersonal & Communication	SVM	118	1.16	.158	1.03	.303
Patient Care	SVM	116	1.16	.162	0.98	.143
Practice-Based Learning	SVM	106	1.29	.125	1.12	.099
Professionalism	SVM	133	1.14	.068	1.21	.023
Medical Knowledge	Lasso	114	1.00	.097	0.96	.020
Systems-Based Practice	ENet	110	1.17	.062	1.19	.018

Note. n = sample size with complete data. RMSE = root mean square error. R^2 = proportion of variance explained. SVM = support vector machine. ENet = elastic net regression. Training metrics from 10-fold cross-validation; test metrics from held-out 20%.

Results showed substantial variability in criterion validity across outcomes. The EVLA constructs demonstrated modest criterion validity for Interpersonal & Communication Skills ($R^2 = .303$, RMSE = 1.03), explaining approximately 30% of the variance in faculty ratings. Weaker but statistically meaningful relationships emerged for patient care performance ($R^2 = .143$, RMSE = 0.98).

In contrast, EVLA constructs showed minimal criterion validity for Practice-based Learning engagement ($R^2 = .099$, RMSE = 1.12), Professionalism ($R^2 = .023$, RMSE = 1.21), Medical Knowledge ($R^2 = .020$, RMSE = 0.96), and Systems-based Practice ($R^2 = .018$, RMSE = 1.19). The average R^2 across all six outcomes was .101, albeit with substantial heterogeneity.

Criterion Validity: Item-Level Analysis

The item-level analysis revealed fundamentally different predictive patterns between clinical performance outcomes and wellbeing/distress outcomes, suggesting domain-specific utility of EVLA measurement granularity. Table 11 presents results from the item-level analysis of clinical performance outcomes, which revealed a strikingly heterogeneous pattern across the six competency domains. Most notably, item-level prediction of Interpersonal & Communication Skills substantially improved. This marked improvement suggests that the specific behavioral indicators captured by individual EVLA items contain meaningful information about Interpersonal & Communication Skills competence that is obscured when items are aggregated into latent factor scores.

Table 11. Best-Performing Model Results for Predicting Clinical Performance Outcomes Using Individual EVLA Items

Outcome	**Model**	**n**	**Train RMSE**	**Train R²**	**Test RMSE**	**Test R²**
Interpersonal & Communication Skills	LM	118	1.28	.190	0.88	.486
Systems-Based Practice	SVM	110	1.20	.096	1.18	.039
Practice-Based Learning	Ridge	106	1.34	.094	1.12	.018
Patient Care	RF	116	1.18	.133	1.01	.010
Professionalism	Lasso	133	1.10	—	1.18	-.000
Medical Knowledge	SVM	114	1.01	.131	0.98	-.169

Note. n = sample size with complete data on all 20 items and outcome. LM = linear regression. RF = random forest. SVM = support vector machine. — indicates metric could not be computed. All models used 20 individual EVLA items as predictors. Training metrics from 10-fold cross-validation; test metrics from held-out 20%.

However, this enhancement was highly outcome specific. Other clinical outcomes showed comparable or substantially worse performance at the item level relative to factor-level prediction. Medical Knowledge, Professionalism, Patient Care, and Practice-Based Learning showed near-zero criterion validity at the item level. Systems-Based Practice showed modest but diminished prediction ($R^2 = .039$) compared to its factor-level performance ($R^2 = .018$).

Lasso regularization identified three specific items as moderately predictive of Interpersonal & Communication Skills performance (Table 12), providing insight into which aspects of learner readiness translate into effective interpersonal functioning in clinical contexts. The two items from the Action construct emerged as the strongest positive predictors: team_per_24 ("Controls anger when dealing with highly stressful situations"; β = .261) and team_per_16 ("Behaves in an ethical manner"; β = -.349).

Table 12. Regularized Regression Coefficients of EVLA Items vs. Core Competency Ratings

			Interpersonal &		**Systems-**	**Practice-**
			Communication	**Patient**	**based**	**based**
Item	**Professionalism**	**Medical Knowledge**	**Skills**	**Care**	**Practice**	**Learning**
team_per_25		-0.113				
team_per_8			0.077	-0.012		
team_per_24			0.261			
self_test_33			-0.017			
team_per_1		0	0.083			
team_per_10		0.08	0.057	0.035		0.145
team_per_16		-0.059	-0.349	-0.112	-0.046	-0.114
team_per_4			0.05			
team_per_7		-0.161		-0.043		-0.227
lasa_2		0.139		0.2		0.085
lasa_4						0.025
lasa_5			0.057			
jspll_5			-0.071	0.023		
jspll_13		-0.045	-0.263			-0.184
jspll_10				-0.216		

Notably, two items showed counterintuitive negative relationships with Interpersonal & Communication Skills performance despite their positive wording and face validity. The Continuity item jspll_13 ("I take every opportunity to gain new knowledge or skills in my job"; β = -.266) and the Capability item team_per_16 ("Behaves in an ethical manner"; β = -.221) both showed substantial negative coefficients. The negative coefficients for jspll_13 and team_per_16 warrant careful interpretation. These findings likely reflect a response bias wherein respondents who strongly endorse universally positive statements about their own behavior may lack the self-awareness or critical self-reflection necessary for accurate self-assessment (Dunning, 2011). Conversely, students who provide more modest self-ratings may demonstrate greater metacognitive accuracy, which itself may be associated with superior actual performance.

Discussion

The present study developed a brief measure of physician capacity based on the Environmentally Valid Learning Approach (EVLA) framework demonstrating good structural validity and strong convergent validity for physician wellbeing outcomes.

The hypothesized five-factor structure of the EVLA demonstrated good fit to the data (CFI = 0.958, RMSEA = 0.052), supporting the theoretical model of physician adaptive capacity as comprising distinct but interrelated domains: Capacity, Capability, Readiness, Action, and Continuity. Cross-validation across 100 random data splits confirmed the stability of this structure (mean CFI = 0.954, SD = 0.018), indicating robust generalizability. Factor loadings were strong for all constructs except Action, which showed marginal reliability (α = 0.68, CR = 0.43) attributable to item heterogeneity.

The theoretical modification adding a covariance between Capacity and Action substantially improved model fit and is conceptually meaningful: foundational wellbeing resources (Capacity) may directly enable emotional regulation processes (Action) independent of the sequential pathway through Capability and Readiness. This finding suggests that physicians with depleted physical, emotional, or social resources will struggle with in-the-moment regulation regardless of their knowledge or preparedness.

The EVLA model demonstrated strong convergent validity through theoretically consistent associations with established wellbeing measures. Capacity showed the strongest correlations across all wellbeing outcomes. Continuity demonstrated moderate correlations with all outcomes. The association with burnout (r = -0.52) suggests a relationship between professional development and burnout, though the causal direction is unclear and possibly a feedback loop. Overall, these findings demonstrate the primacy of wellbeing for physician functioning.

However, given that the Capacity construct is operationalized through four items assessing overall quality of life, physical health, mental health, and emotional wellbeing, there is substantial conceptual overlap with the criterion measures themselves. While this overlap raises concerns about criterion contamination, it simultaneously demonstrates that a brief, four-item assessment can capture the core variance underlying multiple distinct wellbeing measures. This has practical implications: the Capacity subscale may serve as an efficient screening tool that approximates information from longer instruments though at the cost of reduced specificity in identifying specific types of distress.

Criterion Validity

The minimal criterion validity observed between EVLA constructs and supervisor ratings of clinical competence warrants careful interpretation. There are numerous plausible explanations. First, self-report versus observer measurement mismatch may fundamentally limit criterion validity. The EVLA model assesses physicians' subjective perceptions of their capacity, capability, and readiness. In contrast, supervisor ratings capture externally observable behaviors and clinical outputs. These represent distinct constructs that may correlate only modestly even when both are valid.

Second, measurement characteristics of supervisor ratings may attenuate observable relationships. The performance ratings in this sample likely suffer from restricted range, variable rating standards across supervisors, and lack of familiarity with either the target or competency.

Third, EVLA constructs may operate as distal rather than proximal predictors of observable clinical performance. EVLA constructs may predict the potential for effective performance rather than current demonstrated competence, particularly in a sample undergoing remediation.

Theoretical Implications

This study provides the first attempt at empirical validation of the EVLA model, demonstrating that physician adaptive capacity can be conceptualized as five distinct but interrelated domains representing the pathway from foundational resources to sustained professional behavior. The framework received support through the structural model but criterion validity was minimal.

The EVLA model extends beyond traditional competency frameworks by incorporating biopsychosocial foundations (Capacity) and self-reflective processes (Continuity) alongside knowledge/skills (Capability) and performance (Action). This comprehensive approach aligns with calls for ecological models of physician development (Mazmanian et al., 2021) and addresses the critique that medical education overemphasizes knowledge acquisition while neglecting wellbeing and self-awareness (Frith & Frith, 2012).

Clinical and Educational Applications

Needs Assessment for CME/CPD

The EVLA model addresses a critical gap in continuing medical education: the lack of validated tools for assessing learning needs beyond knowledge deficits (Davis et al., 2006). By identifying deficits across five domains, the EVLA model enables targeted intervention tailored to individual needs, assuming proper criterion validity. For example, physicians scoring low on Capacity would benefit from wellness interventions, whereas those scoring low on Continuity would benefit from self-directed learning support and professional development planning to sustain engagement over time. This tailored approach represents a shift from one-size-fits-all continuing education to personalized learning interventions aimed at addressing individual needs.

Early Identification and Intervention

The strong prediction of depression and psychological distress suggests the EVLA model could serve as an early warning system for physicians at risk for serious mental health concerns. The finding that a single item measuring overall emotional health (lasa_4) demonstrates substantial predictive power (given its average coefficient across wellbeing measures) suggests that even briefer screening approaches may be viable. Again, the short but powerful predictiveness of the Capacity measure (comprised of LASA items 1, 2, 4, and 5) could be easily implemented as a short form measure of wellbeing. Furthermore, the link among poor emotional control and lack of communication efficacy with burnout and poor wellbeing suggests these behavioral observations be treated as indicators of distress.

Limitations

Limited Measures of Convergence

The convergent validity testing in this study suffers from a significant imbalance in the availability of theoretically appropriate criterion measures across the five EVLA constructs. Capacity demonstrated the strongest convergent validity evidence; however, we had abundant measures with direct theoretical overlap for Capacity, enabling robust convergent validity testing for this construct.

In stark contrast, the remaining four EVLA constructs lacked theoretically appropriate convergent measures in the available dataset. This gap reflects the novel nature of the EVLA framework: because the model integrates constructs spanning biopsychosocial foundations, professional competence, situational readiness, real-time regulation, and metacognitive monitoring in ways not previously operationalized in physician assessment, existing validated instruments were not designed to map onto these theoretical domains. The field lacks established "gold standard" measures of constructs of such. This gap highlights both the innovation of the EVLA model and the challenge of validating it using existing measurement tools developed for different theoretical frameworks. This asymmetry in construct coverage means that while we can conclude Capacity is validly measured and strongly predicts psychological wellbeing, we have far weaker evidence for the other four constructs.

Criterion Measures and Common Method Variance

The lack of performance measures such as patient outcomes and clinical quality metrics, as well as the limited number of competency ratings limits conclusions about EVLA's utility for predicting actual professional behavior versus self-reported wellbeing. Collateral data rating physician/trainee performance is known to skew high and have limited interrater reliability (Sureda et al., 2021).

Future validation should include such objective performance data. Consensus observer ratings including 360-degree feedback from colleagues would provide stronger external validation. Patient outcomes including satisfaction scores and clinical would test whether EVLA scores translate to patient care quality.

The weak prediction of ACGME competencies in preliminary analyses suggests EVLA may be more relevant for wellbeing than performance per se, though wellbeing and performance are theoretically linked.

Generalizability

Data collection from a single institution limits generalizability in several important ways. The sample was from one institution whose sample of physicians—though referred across the US—should not be assumed to be representative of the US physician population, given the reason for their referral to that site. By the same reasoning, ratings around professionalism and interpersonal functioning should be lower than the general US physician population. Measurement invariance testing could formally examine whether the five-factor structure, factor loadings, and item intercepts remain equivalent across these groups.

Sample Size and Power

While the sample size was adequate for confirmatory factor analysis ($N = 289$), it limited more complex analyses in several ways. Limited ratings precluded more sophisticated machine learning approaches and limited the precision of regression estimates.

Limited Item Scope

A significant limitation of this study concerns the source and selection of items used to operationalize EVLA constructs. Rather than developing items specifically designed to assess the theoretical constructs articulated in the EVLA framework, we adopted a pragmatic approach of mining existing validated instruments for items with face-valid correspondence to our five constructs. Given the inductive development of EVLA and expensive nature of physician self-report data, this was a necessary compromise. This approach, while efficient and ensuring psychometric quality of individual items, introduces several critical problems that constrain interpretation and generalizability.

Important theoretical content may be underrepresented or absent when item pools are constrained by what exists in available instruments rather than what the theory specifies. The EVLA framework identifies the salience of systems-level factors yet there were a limited number of candidates in the item pool to assess this. Similarly, Continuity is theorized to encompass metacognitive monitoring and adaptive self-correction, yet our items predominantly assess engagement and motivation.

Future measurement development must address these limitations through a theoretically driven item development process. This should begin with domain specification, creating detailed content matrices that map each EVLA construct's theoretical components onto specific, observable indicators.

In summary, while the current validation provides preliminary evidence for EVLA's five-factor structure and predictive utility for wellbeing screening, the fundamental measurement limitation of operationalizing a novel theoretical framework using items borrowed from instruments designed for different purposes presents obvious constraints. A purpose-built EVLA instrument, developed through rigorous domain specification and item generation procedures, is requisite to fully effect the framework. Again, performance and competence indicators should also be included to test whether EVLA predicts actual professional behavior rather than only self-reported wellbeing. Relevant criterion measures such as ratings from supervisors and peers, quality of care indicators based on chart review or claims data, patient satisfaction scores from standardized surveys, and social network metrics should be tested against this model.

Conclusion

This study provides the first comprehensive psychometric validation of the Environmentally Valid Learning Approach (EVLA) framework, yielding both promising findings and important limitations that clarify the instrument's scope and appropriate applications. The five-factor structure demonstrated excellent fit and robust cross-validation stability, establishing that physician adaptive capacity can be conceptualized as comprising distinct but interrelated domains spanning foundational biopsychosocial resources (Capacity), professional knowledge and skills (Capability), situation-specific preparedness (Readiness), real-time emotional regulation (Action), and sustained self-reflective engagement (Continuity).

Our findings most clearly establish the Capacity construct for physician psychological health. The robust prediction of burnout, depression, and psychological distress demonstrates that basic physical, emotional, and social resources as a necessary, though not sufficient, condition for sustained professional functioning. Educational programs and healthcare organizations should recognize that interventions targeting knowledge deficits or skill gaps may prove ineffective when physicians lack adequate foundational resources.

However, several critical limitations temper these conclusions and constrain immediate applications. First, the minimal criterion validity for clinical performance outcomes indicates that

EVLA constructs, as currently operationalized, do not reliably predict supervisor ratings of clinical competence. Whether this reflects genuine divergence between self-reported readiness and observable performance, or limitations in our criterion measures remains unclear. Second, purpose-built items developed specifically to assess EVLA's theoretical content are needed to establish whether the five-construct model provides unique insights beyond existing assessment tools. Third, convergent validity testing was adequate only for Capacity, which benefited from abundant wellbeing criterion measures; the other four constructs lacked theoretically matched convergent measures, leaving their validity as distinct constructs inadequately established. Fourth, validation in a remediation sample limits generalizability to practicing physicians without professionalism concerns.

Given these findings, the most defensible current applications of the EVLA instrument are as a wellbeing screener to identify physicians at risk for distress and as a needs assessment for CME programs to identify the presence and nature of existing deficits.

By establishing preliminary psychometric evidence for a comprehensive model of physician adaptive capacity while clearly delineating current limitations and validation gaps, this study provides a foundation for future measurement development and validation research. The EVLA framework's integration of biopsychosocial foundations, professional competence, environmental readiness, emotion regulation, and metacognition represents a rich, holistic approach to understanding physician development beyond competency frameworks.

References

Baumeister, R. F., André, N., Southwick, D. A., & Tice, D. M. (2024). Self-control and limited willpower: Current status of ego depletion theory and research. *Current Opinion in Psychology*, *60*, 101882. https://doi.org/10.1016/j.copsyc.2024.101882

Baumeister, R. F., & Vohs, K. D. (2018). Strength model of self-regulation as limited resource: Assessment, controversies, update. In *Self-regulation and self-control* (pp. 78–128). Routledge.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cruess, R. L., Cruess, S. R., Boudreau, J. D., Snell, L., & Steinert, Y. (2014). Reframing Medical Education to Support Professional Identity Formation: *Academic Medicine*, *89*(11), 1446–1451. https://doi.org/10.1097/ACM.0000000000000427

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of Physician Self-assessment Compared With Observed Measures of Competence: A Systematic Review. *JAMA*, *296*(9), 1094. https://doi.org/10.1001/jama.296.9.1094

Dorman, T., & Miller, B. M. (2011). Continuing Medical Education: The Link Between Physician Learning and Health Care Outcomes: *Academic Medicine*, *86*(11), 1339. https://doi.org/10.1097/ACM.0b013e3182308d49

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Elsevier.

Englander, R., Cameron, T., Ballard, A. J., Dodge, J., Bull, J., & Aschenbrener, C. A. (2013). Toward a Common Taxonomy of Competency Domains for the Health Professions and

Competencies for Physicians: *Academic Medicine*, *88*(8), 1088–1094. https://doi.org/10.1097/ACM.0b013e31829a3b2b

Frank, J. R., Snell, L., & Sherbino, J. (2015). CanMEDS 2015 Physician Competency Framework. *Physician Competency Framework Series I.*

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313.

Golino, H., & Christensen, A. (2025). *EGAnet: Exploratory Graph Analysis – A framework for estimating the number of dimensions in multivariate data using network psychometrics*. https://doi.org/10.32614/CRAN.package.EGAnet

Hojat, M., DeSantis, J., Shannon, S. C., Mortensen, L. H., Speicher, M. R., Bragan, L., LaNoue, M., & Calabrese, L. H. (2018). The Jefferson Scale of Empathy: A nationwide study of measurement properties, underlying components, latent variable structure, and national norms in medical students. *Advances in Health Sciences Education*, *23*(5), 899–920.

Hojat, M., Mangione, S., Nasca, T. J., Cohen, M. J., Gonnella, J. S., Erdmann, J. B., Veloski, J., & Magee, M. (2001). The Jefferson Scale of Physician Empathy: Development and preliminary psychometric data. *Educational and Psychological Measurement*, *61*(2), 349–365.

Hojat, M., Veloski, J., Nasca, T. J., Erdmann, J. B., & Gonnella, J. S. (2006). Assessing Physicians' Orientation Toward Lifelong Learning. *Journal of General Internal Medicine*, *0*(0), 060721075157047-??? https://doi.org/10.1111/j.1525-1497.2006.00500.x

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Irby, D. M., & Hamstra, S. J. (2016). Parting the clouds: Three professionalism frameworks in medical education. *Academic Medicine*, *91*(12), 1606–1611.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kuhn & Max. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Leite, W. L., Huang, I.-C., & Marcoulides, G. A. (2008). Item selection for the development of short forms of scales using an ant colony optimization algorithm. *Multivariate Behavioral Research*, *43*(3), 411–431.

Locke, D. E., Decker, P. A., Sloan, J. A., Brown, P. D., Malec, J. F., Clark, M. M., Rummans, T. A., Ballman, K. V., Schaefer, P. L., & Buckner, J. C. (2007). Validation of single-item linear analog scale assessment of quality of life in neuro-oncology patients. *Journal of Pain and Symptom Management*, *34*(6), 628–638.

Maslach, C., & Jackson, S. E. (1981). The measurement of experienced burnout. *Journal of Organizational Behavior*, *2*(2), 99–113. https://doi.org/10.1002/job.4030020205

Mazmanian, P. E., Cervero, R. M., & Durning, S. J. (2021). Reimagining Physician Development and Lifelong Learning: An Ecological Framework. *Journal of Continuing*

Education in the Health Professions, *41*(4), 291–298. https://doi.org/10.1097/CEH.0000000000000406

Mazmanian, P. E., & Davis, D. A. (2002). Continuing Medical Education and the Physician as a Learner: Guide to the Evidence. *JAMA*, *288*(9), 1057. https://doi.org/10.1001/jama.288.9.1057

Narrow, W. E., Clarke, D. E., Kuramoto, S. J., Kraemer, H. C., Kupfer, D. J., Greiner, L., & Regier, D. A. (2013). DSM-5 Field Trials in the United States and Canada, Part III: Development and Reliability Testing of a Cross-Cutting Symptom Assessment for DSM-5. *American Journal of Psychiatry*, *170*(1), 71–82. https://doi.org/10.1176/appi.ajp.2012.12071000

Peck, J. A., & Porter, T. H. (2022). Pandemics and the Impact on Physician Mental Health: A Systematic Review. *Medical Care Research and Review*, *79*(6), 772–788. https://doi.org/10.1177/10775587221091772

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raborn, A., & Leite, W. (2024). *ShortForm: Automatic Short Form Creation*. https://doi.org/10.32614/CRAN.package.ShortForm

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika*, *34*(S1), 1–97. https://doi.org/10.1007/BF03372160

Schwarzer, R., & Jerusalem, M. (1995). Generalized self-efficacy scale. *J. Weinman, S. Wright, & M. Johnston, Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*, *35*(37), 82–003.

Singh, J. A., Satele, D., Pattabasavaiah, S., Buckner, J. C., & Sloan, J. A. (2014). Normative data and clinically significant effect sizes for single-item numerical linear analogue self-assessment (LASA) scales. *Health and Quality of Life Outcomes*, *12*(1), 187.

Sureda, E., Chacón-Moscoso, S., Sanduvete-Chaves, S., & Sesé, A. (2021). A Training Intervention through a 360° Multisource Feedback Model. *International Journal of Environmental Research and Public Health*, *18*(17), 9137. https://doi.org/10.3390/ijerph18179137

Swiggart, W. H., Williams, M. V., White Williams, B., Dewey, C. M., Ghulyan, M. V., & Wallston, K. A. (2014). ASSESSMENT OF A PHYSICIAN'S WORKPLACE BEHAVIOR. *Physician Leadership Journal*, *1*(2).

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

West, C. P., Dyrbye, L. N., & Shanafelt, T. D. (2018). Physician burnout: Contributors, consequences and solutions. *Journal of Internal Medicine*, *283*(6), 516–529. https://doi.org/10.1111/joim.12752

William Revelle. (2025). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Williams, B. W., Rankin, P., & Williams, M. V. (2016). Understanding disruptive behavior in the seasoned clinician. *Physician Leadership Journal*, *3*(6), 58.

Williams, B. W., Welindt, D., Hafferty, F. W., Stumps, A., Flanders, P., & Williams, M. V. (2021). Adverse Childhood Experiences in Trainees and Physicians With Professionalism Lapses: Implications for Medical Education and Remediation. *Academic Medicine*, *96*(5), 736–743. https://doi.org/10.1097/ACM.0000000000003532

Williams, B. W., & Williams, M. V. (2020). Understanding and Remediating Lapses in Professionalism: Lessons From the Island of Last Resort. *The Annals of Thoracic Surgery*, *109*(2), 317–324. https://doi.org/10.1016/j.athoracsur.2019.07.036

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17. https://doi.org/10.18637/jss.v077.i01