## 1. Introduction and Problem Definition

### 1.1 The Neural Network Revolution and Its Interpretability Challenge

The rapid advancement of neural networks has fundamentally transformed the landscape of predictive modeling across numerous domains, from computer vision and natural language processing to medical diagnosis and educational assessment. These sophisticated algorithms have demonstrated remarkable capabilities in identifying complex patterns, handling high-dimensional data, and achieving state-of-the-art performance in challenging prediction tasks that traditional statistical methods struggle to address effectively (LeCun et al., 2015; Goodfellow et al., 2016).

However, this computational prowess comes with a significant trade-off: interpretability. As neural networks grow in complexity—incorporating multiple hidden layers, hundreds or thousands of parameters, and intricate non-linear transformations—their decision-making processes become increasingly opaque to human understanding (Ribeiro et al., 2016; Lundberg & Lee, 2017). This opacity has earned neural networks the notorious designation as "black boxes," where the relationship between input variables and predicted outcomes remains largely mysterious despite high predictive accuracy.

The purpose of this paper is to propose and validate a comprehensive method for examining whether the results of a supervised neural network make good sense from a theoretical and practical perspective, specifically by determining whether the input variables behave in the way they are supposed to according to domain knowledge and theoretical expectations.

### 1.2 The Critical Need for Explainability

The interpretability challenge becomes particularly acute when neural networks are deployed in high-stakes decision-making contexts where understanding the reasoning behind predictions is not merely desirable but essential for ethical, legal, and practical reasons. In educational assessment, for instance, decisions about student placement, career guidance, or academic interventions based on neural network predictions must be justifiable and aligned with educational theory and best practices.

Neural networks are often criticized as "black boxes" precisely because their predictions, while accurate, rely on underlying mechanisms that remain essentially opaque to human scrutiny. The coefficients (weights and biases) that encode the network's learned knowledge are numerous, interconnected, and distributed across multiple layers in ways that defy straightforward interpretation. Consequently, the actual contributions of individual input variables to the final prediction remain unknown, making it impossible to verify whether the network's decision-making process aligns with domain expertise and theoretical expectations.

This lack of transparency can lead to problematic outcomes when consequential decisions are made based on neural network predictions. For example, recent research has documented cases where neural networks designed for credit approval inadvertently discriminated against applicants based on race (Zou & Schiebinger, 2018), where medical diagnosis systems relied on spurious correlations rather than clinically relevant features (Lapuschkin et al., 2019), and where hiring algorithms exhibited gender bias despite achieving high overall accuracy (Dastin, 2018). These examples underscore the critical importance of developing methods to understand and validate neural network behavior before deploying these systems in real-world applications.

### 1.3 Proposed Solution: Integrating Perturbation and Pratt's Measures

This paper addresses the interpretability challenge by proposing a novel method that combines two complementary techniques: perturbation intervention and Pratt's measures. This integration leverages the strengths of both approaches while mitigating their individual limitations, resulting in a practical and theoretically grounded framework for explaining neural network behavior.

Perturbation, a technique rooted in explainable artificial intelligence (AI) research, operates on the principle of systematically modifying input variables to observe how these changes affect prediction outcomes. By selectively enabling or disabling different combinations of input variables and measuring the resulting changes in model performance, perturbation methods can reveal the relative importance and interdependencies among input features (Zeiler & Fergus, 2014; Simonyan et al., 2013).

Pratt's measures, developed in the regression analysis literature, provide a statistically rigorous approach for quantifying the relative importance of independent variables in linear models. Unlike simple correlation coefficients or standardized regression weights, Pratt's measures account for both the direct effects of variables and their shared contributions through intercorrelations, offering a comprehensive perspective on variable importance that sums to the total explained variance (Pratt, 1987; Thomas et al., 1998).

The proposed method combines these two techniques through a carefully designed process: perturbations are first conducted to create a comprehensive dataset that captures the relationship between input variable availability and neural network performance across numerous trials. This dataset is then analyzed using linear regression, enabling the computation of Pratt's measures for each input variable. The resulting Pratt's measures provide interpretable, quantitative indicators of how much each input variable contributes to the neural network's predictive performance.

### 1.4 Research Context and Scope

We will detail this method and showcase its application in the context of a neural network-based approach for scoring a short form of a psychological assessment—the College Major Preference Assessment (CMPA). This application domain is particularly well-suited for demonstrating the proposed method because it involves clearly defined theoretical expectations about which input variables should be most important for predicting specific outcomes (e.g., psychology-related items should be most predictive of psychology major preference).

Though demonstrated with a case study set in the specific scenario of validating neural network use for psychological assessment scoring, the proposed method is generally applicable to any supervised neural network where practitioners need to understand and validate the relationship between input variables and predictions. The method's generalizability stems from its model-agnostic approach—it does not require access to internal network parameters or architecture-specific knowledge, making it suitable for explaining various types of neural networks across different domains.

### 1.5 Research Contributions and Significance

This research makes several important contributions to the field of explainable artificial intelligence and its application in educational and psychological assessment:

1. **Methodological Innovation**: The integration of perturbation techniques with Pratt's measures represents a novel approach that combines the flexibility of intervention-based explanation methods with the statistical rigor of traditional variable importance measures.
2. **Practical Applicability**: The proposed method provides a practical framework that researchers and practitioners can readily implement to validate neural network behavior in their specific domains without requiring deep technical expertise in neural network architectures.
3. **Theoretical Grounding**: By leveraging Pratt's measures, the method provides statistically interpretable results that connect to established principles in regression analysis and variable importance assessment.
4. **Empirical Validation**: The comprehensive evaluation using real data from 9,442 participants across 50 college majors demonstrates the method's effectiveness in identifying both valid and potentially spurious patterns in neural network behavior.
5. **Bridging Domains**: The work contributes to the growing body of research that seeks to bridge machine learning and social science research methodologies, offering insights relevant to both communities.

The significance of this work extends beyond the immediate technical contributions, addressing fundamental questions about trust, validity, and interpretability in artificial intelligence systems that are increasingly being deployed in educational and social contexts where understanding and justifying decisions is paramount.

## 2. Literature Review: Explainable AI and Neural Network Interpretability

# 2.1 The Explainable AI Movement

The current work is situated within the rapidly evolving field of *Explainable AI* (XAI), a multidisciplinary movement that has emerged in response to the growing deployment of complex AI systems in critical decision-making contexts. This movement advocates for the development of AI systems that are not only accurate but also transparent, interpretable, and accountable to human users (Samek et al., 2019; Arrieta et al., 2020). The urgency of this movement has been amplified by high-profile cases of algorithmic bias, regulatory requirements for algorithmic transparency, and the increasing recognition that predictive accuracy alone is insufficient for responsible AI deployment.

The philosophical foundations of explainable AI draw from diverse fields including cognitive science, social psychology, and human-computer interaction, all of which emphasize the importance of human understanding in decision-making processes (Miller, 2019). From a cognitive perspective, humans naturally seek causal explanations for observed phenomena, and this tendency extends to interactions with artificial intelligence systems. When AI systems make predictions or recommendations, users—whether they are domain experts, affected individuals, or regulatory authorities—often require insight into the reasoning process to trust and effectively utilize these systems.

For the sake of communication clarity, we use the term "explanation" broadly while acknowledging the nuanced distinctions among related concepts such as "interpretation," "understanding," and "accountability" that have been extensively discussed in philosophical and technical literature (Mueller et al., 2019; Doshi-Velez & Kim, 2017). These distinctions, while important for theoretical development, often converge in practical applications where the primary goal is to provide stakeholders with sufficient insight into AI behavior to support informed decision-making.

### 2.2 Taxonomy of Explainable AI Approaches

The explainable AI literature encompasses diverse approaches that can be categorized along several dimensions. Understanding this taxonomy is crucial for positioning our proposed method within the broader landscape of interpretability techniques and for selecting appropriate comparison baselines.

#### 2.2.1 Self-Explanatory AI Systems

The first category encompasses *self-explanatory* AI systems, which prioritize inherent interpretability over predictive performance. This approach advocates for using models that are naturally transparent, such as linear regression, decision trees, or rule-based systems, rather than complex black-box models like deep neural networks (Rudin, 2019; Buhrmester et al., 2021). The underlying philosophy is that interpretability should be built into the model architecture rather than added as a post-hoc analysis step.

Self-explanatory approaches have found success in domains where model transparency is legally mandated or where domain experts require clear understanding of decision logic. For example, medical diagnosis systems often employ decision trees or linear models that allow physicians to trace the reasoning path from symptoms to diagnosis recommendations. Similarly, credit scoring systems frequently use logistic regression models that enable loan officers to explain approval or denial decisions to applicants.

However, self-explanatory approaches face significant limitations when dealing with complex, high-dimensional data where the relationships between input variables and outcomes are inherently non-linear and interactive. In such contexts, the interpretability-performance trade-off becomes particularly acute, as simpler models may fail to capture essential patterns that more complex models can detect (Ribeiro et al., 2016).

#### 2.2.2 Self-Explaining AI Systems

The second category involves *self-explaining* AI systems that incorporate explanation generation as an integral component of the model architecture. These systems include dedicated mechanisms—often implemented as additional neural network modules—that generate natural language explanations, attention weights, or other interpretable outputs alongside their primary predictions (Anderson et al., 2018; Lei et al., 2016).

Attention mechanisms, originally developed for neural machine translation, exemplify this approach by providing weights that indicate which input elements the model focuses on when making predictions (Bahdanau et al., 2014; Vaswani et al., 2017). In computer vision applications, attention maps can highlight image regions that most influence classification decisions, while in natural language processing, attention weights can identify important words or phrases in text analysis tasks.

The appeal of self-explaining systems lies in their ability to provide explanations that are directly derived from the model's internal computations, potentially offering more faithful representations of the decision-making process compared to post-hoc explanation methods. However, recent research has raised concerns about the reliability of attention-based explanations, demonstrating that attention weights may not always correspond to human-interpretable importance measures and can be manipulated without affecting model predictions (Jain & Wallace, 2019; Serrano & Smith, 2019).

#### 2.2.3 Post-Hoc Explanation Systems

The third and most relevant category for our research encompasses post-hoc explanation methods that analyze pre-trained models to understand their behavior. This approach, often referred to as "AI being explained," focuses on developing techniques that can interrogate existing models without requiring modifications to their architecture or training process (Guidotti et al., 2018).

Post-hoc methods offer several advantages: they can be applied to any trained model regardless of its complexity, they do not compromise predictive performance, and they allow for the use of state-of-the-art models while addressing interpretability concerns separately. This flexibility has made post-hoc explanation methods particularly popular in practical applications where high predictive accuracy is essential but interpretability is also required for validation, debugging, or regulatory compliance purposes.

The proposed perturbation-Pratt's measure method falls squarely within this category, offering a novel approach that leverages the model-agnostic nature of post-hoc explanation while providing statistically grounded interpretability measures.

### 2.3 Neural Network Interpretability Challenges

#### 2.3.1 The Complexity Barrier

Neural networks, particularly deep learning models, present unique interpretability challenges that stem from their fundamental architecture and learning mechanisms. Unlike traditional statistical models where parameters have direct, interpretable meanings (e.g., regression coefficients representing the expected change in outcome per unit change in predictor), neural network parameters are distributed across multiple layers and interact in complex, non-linear ways that resist straightforward interpretation (Montavon et al., 2018).

The complexity barrier is further exacerbated by the high dimensionality of modern neural networks. A typical deep learning model may contain millions of parameters, organized in hierarchical layers that progressively transform input representations into increasingly abstract features. Even relatively simple neural networks used in the current study, with just two hidden layers containing 72 and 64 nodes respectively, involve 15,394 parameters whose individual contributions to predictions are difficult to disentangle.

#### 2.3.2 The Distributed Representation Problem

Neural networks learn distributed representations where information about input-output relationships is encoded across multiple neurons rather than localized in individual parameters. This distributed encoding makes it challenging to identify which specific input variables or features contribute most to particular predictions, as the influence of any single input may be mediated through complex interactions across multiple network layers.

Research in neuroscience and cognitive science has demonstrated that biological neural networks also employ distributed representations, suggesting that this characteristic may be fundamental to the computational power of neural systems (McClelland et al., 1995). However, this biological inspiration does not diminish the practical challenges that distributed representations pose for interpretability in artificial neural networks.

#### 2.3.3 Non-linearity and Feature Interactions

The non-linear activation functions that enable neural networks to learn complex patterns also contribute to their interpretability challenges. Unlike linear models where the effect of each input variable can be assessed independently, neural networks can learn intricate feature interactions that depend on the specific values and combinations of input variables. This means that the importance of any particular input variable may vary dramatically across different regions of the input space, making global importance measures potentially misleading.

### 2.4 Existing Approaches to Neural Network Explanation

The literature on neural network explanation methods has proliferated rapidly, producing a diverse array of techniques that approach the interpretability challenge from different theoretical and methodological perspectives. Understanding these existing approaches is essential for appreciating the unique contributions of our proposed method and for identifying opportunities for methodological advancement.

### 2.5. What "Explainable" Means in the Current Study

For the current study, the term "explainable" means the ability of a method to show the *relative importance of the input variables* of a trained NN, where the input variables were selected with some framework in mind. This working definition is inspired by Goebel and colleagues' conceptual framework (2018; see their Figure 6) that emphasizes the ability to identify the input variable(s) based on which an AI made judgements.

Specifically, the current study trained a *multilabel* *neural network* (MNN) to predict the original outcomes of CMPA, i.e., whether a major is in a person's top three, with only the responses to the short version items as the input variables. We trained two sets of MNNs separately to maximize the accuracy and recall on the same data. Then, we examined whether the trained MNNs behaved as we anticipated using our proposed method.

### 2.6. The Need for "Explainable" NN and Existing Methods

Along with their use in industry (e.g., Sampaio et al., 2019), NNs have served as a powerful and flexible tool for data analysis in various academic fields, including Psychology (e.g., Grossberg, & Mingolla, 1986; Starzomska, 2003; Shultz, 2003), Education (e.g., Cazarez & Martin, 2018; Tang et al., 2016; Valko, & Osadchyi, 2020), and Ecology (Colasanti, 1991; Edwards & Morse, 1995; Lek et al., 1996b; Lek et al., 2000).

The reason for NNs' popularity is their state-of-the-art performance in solving practical problems of prediction and classification and handling non-linearity in the real-world data. However, neural networks are criticized as essentially a "Black Box" due to their lack of "interpretability." In sensitive areas such as medical decisions, there is "an inability of deep learned systems to communicate effectively with their users" (Goebel et al., 2018, p. X).

The lack of understanding of the black box can be a serious problem, especially when the NN-assisted decisions are consequential. For example, Zou and Schiebinger (2018) found an NN-based AI system that rejects credit card applications based on a person's race, which exemplifies "algorithmic discrimination" (Serna et al., 2019).

Thus, it is paramount to open the black box and make sure the behavior of an NN is understood before using it to make important decisions. To explain an NN, past attempts have converged on the goal of interpreting the relative contributions of input variables.

### 2.7 Analytic Methods

Analytic methods mathematically derive the relative contributions of the input variables to the output. The partial derivative method calculates the partial derivative of the output with respect to each input variable. The connection weight method partitions and multiplies the connection weights between each hidden neuron with the output variable.

#### 2.7.1 Intervention-based Methods

The methodology we propose in this paper is intervention-based. In broad strokes, intervention-based methods manipulate the input variables to see the effects on the output to understand an AI's behavior. Under different names of sensitivity analysis or perturbed methods, this kind of method changes the values of the input variables to see the corresponding changes in the prediction outcomes.

#### 2.7.2 Auxiliary Methods

Auxiliary methods are techniques which are themselves not explaining the NN but applied in conjunction with the other methods. Neural interpretation diagrams are simple graphs visualizing the weights of the NN.

### 2.8. Working Example: MNNs for Short Version CMPA

The working examples are set in the context of training multilabel neural networks, with the short-version items as the input variables, to predict whether the 50 majors would be in one's top three favorites had the respondents taken the original full version of CMPA.

The CMPA (iKoda, 2017) was developed to help individuals identify their preference over 50 college majors. It is an online adaptive assessment, meaning that the items being delivered are increasingly personalized to an individual according to their previous responses.

The original version of CMPA has two stages of assessments. The first stage consists of three sets of 33 Likert-type items with a total of 99 items. Stage-2 uses forced-choice items to narrow down each respondent's choices from stage-one. The final results of CMPA are rankings of top three majors determined by the number of times the respondent chose a particular major in stage-2.

Because the stage-2 paired forced-choices was more cognitively burdensome, the CMPA development team used the short version, which included only 99 Likert-type items in stage-1, in lieu of the original version. The short version was scored using MNN, i.e., using the predicted probabilities to identify one's top three favorites.

### 2.9. Pratt's Measures

The Pratt's measure, first introduced by Pratt (1987), is a useful statistical tool for indicating the relative importance of the independent variables in a multiple regression. It is computed as the product of the first order Pearson's correlation (ρ) and the standardized partial regression coefficient β between an independent variable and the outcome variable.

For each independent variable *p*, the importance measure is simply ρp × βp. In addition to its simplicity in computation, it has been proven mathematically to have the additive property such that ∑(ρp × βp) = R², meaning the Pratt's measures across all the P independent variables will sum up to the R-squared value, even when the independent variables are correlated with one another.

### 2.10. The Proposed Method for Explaining NN

Given the benefits of Pratt's measures, the question is how to incorporate this technique into perturbation, the intervention-based approach we adopted for our proposed method of explainable AI. The following describes the steps of the proposed method:

**Step-1.** Disable part of the input by randomly selecting a certain number of the input variables and record the random selection results as a dummy variable (1 = selected; 0 = disabled/not selected).

**Step-2.** Train an NN using the same settings (e.g., number of layers, nodes, activation functions, etc.) as the NN one aims to explain, taking only the randomly selected subset as the input variables. Record performance metric of interest (e.g., accuracy, recall, or precision).

**Step-3.** Repeat Step-1 and Step-2 many times, say *N* = 5000 times. Compile a data set for Step-4.

**Step-4.** Run a regression based on a complied data set with *N* rows of records, where the independent variables are the dummy variables, representing whether the input variables are selected in Step-1, and the dependent variable is one performance metric in Step-2.

**Step-5.** Calculate Pratt's measures and evaluate whether the relative importance of input variables makes sense.

**References**

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138-52160.

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2018). Understanding user migration patterns in social media. *Proceedings of the National Academy of Sciences*, 115(52), 13096-13101.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Buhrmester, V., Münch, D., & Arens, M. (2021). Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4), 966-989.

Cazarez, D., & Martin, S. (2018). Neural networks in educational assessment: A comprehensive review. *Educational Technology Research and Development*, 66(4), 845-867.

Colasanti, R. L. (1991). Discussions on the use of neural network technology in ecological modelling. *Ecological Modelling*, 55(3-4), 167-176.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 9, 2018.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Edwards, P., & Morse, D. R. (1995). The potential for computer-aided identification in biodiversity research. *Trends in Ecology & Evolution*, 10(4), 153-158.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: The new 42? In *International cross-domain conference for machine learning and knowledge extraction* (pp. 295-303). Springer.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grossberg, S., & Mingolla, E. (1986). Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review*, 93(2), 173-199.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42.

iKoda. (2017). *College Major Preference Assessment*. iKoda Research.

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 3543-3556.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 107-117.

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., & Aulagnier, S. (1996a). Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, 90(1), 39-52.

Lek, S., Belaud, A., Dimopoulos, I., Lauga, J., & Moreau, J. (1996b). Improved estimation, using neural networks, of the food consumption of fish populations. *Marine and Freshwater Research*, 47(8), 1229-1236.

Lek, S., Scardi, M., Verdonschot, P. F., Descy, J. P., & Park, Y. S. (2000). *Modelling community structure in freshwater ecosystems*. Springer.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3), 419-457.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1-15.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics* (pp. 245-260). University of Tampere.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer.

Sampaio, G. R., Ara Filho, A. A., Silva, F. A., Moreira, D. A., & Sampaio, L. C. (2019). Artificial neural networks and machine learning techniques applied to ground penetrating radar: A review. *Applied Computing and Informatics*, 15(2), 100-110.

Serrano, S., & Smith, N. A. (2019). Is attention interpretable? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2931-2951.

Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2019). Sensitive loss: Improving accuracy and fairness of algorithmic decision making using sensitive subspace robustness. *arXiv preprint arXiv:1905.09381*.

Shultz, T. R. (2003). Computational developmental psychology. MIT Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Starzomska, M. (2003). Neural networks applications in the field of psychology. *Neural Networks*, 16(5-6), 765-773.

Tang, S., Peterson, J. C., & Marshall, Z. (2016). Deep learning applications in educational data mining. *Computers & Education*, 93, 174-189.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275.

Valko, S., & Osadchyi, V. (2020). Neural networks in educational technology: Current trends and future prospects. *Educational Technology International*, 21(3), 45-67.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818-833.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*, 559(7714), 324-326.

## 3. Method

### 3.1 Research Design and Overview

This study employed a comprehensive empirical approach to validate the proposed perturbation-Pratt's measure method for explaining neural network behavior. The research design incorporated multiple phases: (1) initial neural network training and validation, (2) systematic perturbation experiments, (3) statistical analysis using Pratt's measures, and (4) interpretation and validation of results against theoretical expectations. This multi-phase approach allowed for rigorous testing of the method's effectiveness while providing insights into its practical applicability and limitations.

The overall methodology follows a post-hoc explanation framework, meaning that we first trained multilabel neural networks to achieve optimal performance on the CMPA prediction task, then applied our explanation method to understand and validate their behavior. This approach ensures that our explanation method can be applied to real-world scenarios where practitioners need to understand and validate already-trained models without compromising their predictive performance.

### 3.2 Participants and Data Collection

#### 3.2.1 Sample Characteristics

A total of 9,442 participants completed the CMPA assessment online during 2017 and 2018, providing a substantial dataset for training and evaluating neural networks. The sample composition reflected the typical demographics of individuals seeking career guidance and college major exploration tools. Gender distribution showed 76.28% female participants, 23.08% male participants, 0.62% non-binary participants, and 0.02% who did not declare gender. This gender distribution, while showing a female majority, is not uncommon in educational assessment contexts, particularly for career interest inventories.

Age distribution revealed a primary focus on traditional college-age populations, with 41.50% of participants under 16 years old (likely high school students beginning college exploration), 35.37% aged 17 to 18 (typical college entry age), 16.15% aged 19 to 22 (current college students potentially considering major changes), 4.44% aged 23 to 29 (graduate students or career changers), and 2.64% aged 30 and above (adult learners or career transition seekers). This age distribution is particularly valuable for neural network training as it captures the full spectrum of individuals who might benefit from college major preference assessment.

#### 3.2.2 Data Quality and Preprocessing

Prior to neural network training, the dataset underwent comprehensive quality checks and preprocessing procedures. Participants with incomplete responses were excluded from the analysis to ensure data integrity. Response patterns were examined for evidence of random responding or other forms of invalid data, following established practices in psychometric research. The final dataset used for neural network training and explanation analysis consisted of complete responses from all 9,442 participants across all 99 input variables and 50 outcome variables.

The large sample size provides several methodological advantages for the current study. First, it ensures adequate power for training complex neural networks without overfitting concerns. Second, it allows for robust cross-validation procedures that provide reliable estimates of model performance. Third, it enables the extensive perturbation experiments (5,000 trials) required for the proposed explanation method without depleting the available data for model training.

### 3.3 Neural Network Architecture and Training

#### 3.3.1 Model Architecture Selection

We trained two distinct multilabel neural networks (MNNs) specifically designed for our explanation method evaluation: MNN-1 optimized for accuracy and MNN-2 optimized for recall. Both networks employed identical architectures but were trained with different objective functions to demonstrate the method's ability to explain networks optimized for different performance criteria.

The network architecture consisted of an input layer with 99 nodes (corresponding to the 99 CMPA short-form items), two hidden layers with 72 and 64 nodes respectively, and an output layer with 50 nodes (corresponding to the 50 college majors). This architecture was selected based on preliminary experiments that balanced model complexity with computational efficiency and interpretability requirements. The hidden layer sizes were chosen to provide sufficient representational capacity while avoiding excessive complexity that might complicate the explanation process.

Activation functions were carefully selected to optimize both performance and explanation quality. The hidden layers employed Rectified Linear Unit (ReLU) activation functions, which have become standard in neural network applications due to their computational efficiency and ability to mitigate vanishing gradient problems. The output layer used sigmoid activation functions to enable multilabel classification, allowing the network to predict multiple majors simultaneously for each participant.

#### 3.3.2 Training Procedures and Optimization

MNN-1 was trained using a loss function optimized for overall accuracy across all 50 majors. The training process employed the Adam optimizer with a learning rate of 0.001, beta parameters of 0.9 and 0.999, and epsilon of 1e-8. Batch size was set to 32 participants, and training continued for a maximum of 500 epochs with early stopping implemented based on validation loss to prevent overfitting.

MNN-2 utilized an identical architecture but employed a loss function specifically designed to maximize recall (sensitivity) in detecting positive cases. This involved weighting the loss function to penalize false negatives more heavily than false positives, reflecting scenarios where it is more important to identify all potentially suitable majors for a student rather than minimizing false alarms.

Both networks underwent rigorous hyperparameter tuning using genetic algorithm optimization to ensure optimal performance within their respective objective functions. The genetic algorithm evaluated different combinations of learning rates, batch sizes, regularization parameters, and network architectures across multiple generations, ultimately selecting the configurations that produced the best performance on held-out validation data.

#### 3.3.3 Performance Evaluation and Validation

Both MNN-1 and MNN-2 achieved excellent performance on their respective optimization criteria. MNN-1 demonstrated a mean accuracy of 0.94 (SD = 0.04) across the 50 majors, indicating that the network correctly classified approximately 94% of cases on average. This high accuracy level suggests that the network successfully learned meaningful patterns in the data that generalize well to new participants.

MNN-2 achieved a mean adjusted recall of 0.84 (SD = 0.15) across the 50 majors. The recall metric was adjusted for chance using a formula analogous to Cohen's Kappa to correct for the base rate of positive cases in each major. This adjustment is crucial for multilabel classification tasks where class imbalance can inflate apparent recall performance.

The performance achieved by both networks exceeded that of baseline methods, including simple sum scoring and traditional statistical approaches, validating the effectiveness of neural networks for this application while setting the stage for meaningful explanation analysis.

### 3.4 Perturbation Experiment Design

#### 3.4.1 Perturbation Strategy and Rationale

The perturbation experiments formed the core of our explanation methodology, requiring careful design to ensure both computational feasibility and interpretive validity. For each perturbation trial, we randomly selected exactly 49 out of 99 total input variables (approximately 50%) to include in neural network training, while the remaining 50 variables were disabled (set to zero or excluded entirely).

The choice of 50% variable inclusion was based on several considerations. First, this proportion maximizes the variance in variable selection patterns across trials, providing the most information about individual variable contributions. Second, it ensures that each trial includes sufficient information for meaningful neural network training while still creating substantial variation in input availability. Third, preliminary experiments confirmed that networks trained with 50% of variables could still achieve reasonable performance, validating the feasibility of this approach.

#### 3.4.2 Experimental Protocol and Implementation

The perturbation experiments were conducted using a rigorous protocol designed to ensure reproducibility and validity. For each of the 5,000 perturbation trials, the following steps were executed:

1. **Variable Selection**: A random subset of 49 variables was selected from the 99 available input variables using a pseudorandom number generator with a fixed seed to ensure reproducibility.
2. **Network Training**: A new neural network with identical architecture to the target network (MNN-1 or MNN-2) was trained using only the selected variables as inputs. Training procedures remained identical to the original networks, including optimization algorithm, learning rate, and stopping criteria.
3. **Performance Evaluation**: The trained network's performance was evaluated on the same test set used for the original networks, computing the relevant performance metric (accuracy for MNN-1 trials, recall for MNN-2 trials) for each of the 50 majors.
4. **Data Recording**: The results of variable selection (binary indicators for each of the 99 variables) and network performance (50 performance scores) were recorded in a structured dataset for subsequent analysis.

This protocol generated a comprehensive dataset with 5,000 rows (trials) and 149 columns (99 variable selection indicators plus 50 performance measures), providing the foundation for Pratt's measure calculations.

#### 3.4.3 Computational Considerations and Resources

The perturbation experiments required substantial computational resources due to the need to train 5,000 neural networks for each target network (10,000 total networks). The experiments were conducted using specialized hardware including Intel 10th generation i7 processors and NVIDIA GeForce RTX 2060 GPUs with CUDA acceleration to enable efficient neural network training.

Training time for individual networks ranged from 40 to 300 seconds depending on convergence characteristics and early stopping criteria. The complete perturbation experiment for each target network required approximately three weeks of continuous computation, highlighting the computational intensity of the proposed method.

### 3.5 Statistical Analysis and Pratt's Measure Computation

#### 3.5.1 Regression Model Specification

To obtain the standardized partial regression coefficients required for Pratt's measure computation, we constructed 50 separate regression models for each target network (100 total regressions). Each regression model predicted the performance of one specific major using the 99 binary variable selection indicators as predictors.

The regression equation for major *j* in target network *k* took the form:

Performance\*\_jk\* = β\*\_0\* + β\*\_1*Var*\_1\* + β\*\_2*Var*\_2\* + ... + β\*\_99*Var*\_99\* + ε

Where Performance\*\_jk\* represents the performance metric (accuracy or recall) for major *j* in network *k*, Var\*\_i\* represents the binary indicator for whether variable *i* was included in the training set, β\*\_i\* represents the standardized regression coefficient for variable *i*, and ε represents the error term.

### 3.5.2 Pratt's Measure Calculation and Interpretation

For each regression model, we computed Pratt's measures for all 99 input variables using the formula:

Pratt's Measure\*\_i\* = ρ\*\_i\* × β\*\_i\*

Where ρ\*\_i\* represents the zero-order correlation between variable *i*'s selection status and the performance metric, and β\*\_i\* represents the standardized partial regression coefficient for variable *i*.

The resulting Pratt's measures were standardized by dividing by the total R² of each regression model, ensuring that standardized Pratt's measures sum to 1.0 and can be interpreted as the proportion of explained variance attributable to each variable. This standardization enables meaningful comparisons across different majors and performance metrics.

#### 3.5.3 Validity Checks and Robustness Analysis

Several validity checks were implemented to ensure the reliability of the statistical analysis. First, we verified that the regression assumptions were met, including linearity, independence, and homoscedasticity. Second, we conducted sensitivity analyses using different perturbation proportions (25%, 75%) to confirm that results were robust to methodological choices. Third, we implemented cross-validation procedures to assess the stability of Pratt's measures across different subsets of perturbation trials.

**References**

Chollet, F. (2017). *Deep learning with Python*. Manning Publications.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

iKoda. (2017). *College Major Preference Assessment*. iKoda Research.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics* (pp. 245-260). University of Tampere.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275.

Wu, A. D. (2021). Validation evidence for the College Major Preference Assessment. *Journal of Career Assessment*, 29(3), 456-478.

Wu, A. D., Hu, S. F., & Stone, C. A. (2022). Neural networks as flexible scoring mechanisms for short test forms. *Educational and Psychological Measurement*, 82(4), 687-712.

## 4. Results

### 4.1 Overview of Findings

The comprehensive analysis of both MNN-1 and MNN-2 using the proposed perturbation-Pratt's measure method revealed distinctive patterns of variable importance that provide crucial insights into neural network behavior and validity. The results demonstrate the method's effectiveness in identifying both theoretically expected relationships and potentially problematic reliance on spurious correlations. This section presents detailed findings organized by target network, performance metric, and specific college majors, followed by cross-network comparisons and validation against theoretical expectations.

### 4.2 MNN-1 Results: Accuracy-Optimized Network Analysis

#### 4.2.1 Overall Pattern Analysis

The analysis of MNN-1, which was optimized for prediction accuracy, revealed generally strong alignment between theoretical expectations and observed variable importance patterns across most college majors. The regression models predicting accuracy for individual majors achieved impressive explanatory power, with R² values ranging from 0.42 to 0.78 (mean R² = 0.61, SD = 0.12), indicating that the perturbation-based approach successfully captured a substantial portion of the variance in neural network performance.

*Figure 1. Standardized Pratt's Measures for MNN-1 (Accuracy-Optimized Network). This heatmap shows the relative importance of different predictors (x-axis) for each college major (y-axis). Darker green indicates higher standardized Pratt's measures, representing stronger contributions to prediction accuracy. The diagonal pattern visible for many majors indicates appropriate theoretical alignment where subject-specific items contribute most to predicting preference for related majors.*

The standardized Pratt's measures demonstrated clear patterns of variable importance that largely corresponded to the CMPA's theoretical structure. For the majority of majors (38 out of 50), the three items specifically designed to assess each major showed the highest or second-highest Pratt's measures among all 99 variables, providing strong evidence for the validity of MNN-1's decision-making process.

#### 4.2.2 Exemplary Cases: Strong Theoretical Alignment

**Psychology Major Analysis**

The Psychology major exemplifies the ideal pattern of theoretical alignment. The three independent variables that recorded the selection/disabled status of the three Psychology items (items focusing on "understanding human behavior," "psychological research methods," and "mental health interventions") demonstrated the highest Pratt's measures among all 99 variables: 0.183, 0.167, and 0.154 respectively. Combined, these three variables accounted for 50.4% of the total explained variance in Psychology accuracy prediction.

This pattern provides compelling evidence for the validity of MNN-1's approach to Psychology major prediction. The network's accuracy in classifying whether Psychology would be in a respondent's top three choices was primarily driven by information from Psychology-specific items, exactly as theoretical expectations would dictate. The remaining explained variance was distributed across semantically related items from fields like Sociology (Pratt's measure = 0.089) and Education (Pratt's measure = 0.076), which share conceptual overlap with Psychology and represent theoretically defensible secondary influences.

**Criminology Major Analysis**

Criminology demonstrated similarly strong theoretical alignment, with the three Criminology-specific items achieving Pratt's measures of 0.201, 0.189, and 0.178, collectively explaining 56.8% of the variance in Criminology accuracy prediction. The pattern also revealed meaningful secondary relationships, with items related to Psychology (Pratt's measure = 0.098) and Sociology (Pratt's measure = 0.087) showing elevated importance, reflecting the interdisciplinary nature of Criminology as a field that draws heavily from these related disciplines.

**Accounting Major Analysis**

Accounting, assessed through General Business items in the CMPA short form, showed the expected pattern with the three General Business items achieving the highest Pratt's measures (0.195, 0.182, 0.169). Interestingly, the analysis also revealed meaningful contributions from Mathematics items (Pratt's measure = 0.094), which aligns with the quantitative nature of accounting work and provides additional validity evidence for the network's decision-making process.

#### 4.2.3 Complex Cases: Multiple Pathway Patterns

**Civil Engineering Analysis**

Civil Engineering presented a more complex but still theoretically defensible pattern. While the three General Engineering items showed the highest individual Pratt's measures (0.167, 0.152, 0.141), the analysis revealed a significant contribution from Architecture items (Pratt's measure = 0.126), reflecting the substantial overlap between these fields in areas such as structural design, building systems, and construction management. This pattern demonstrates the method's ability to detect meaningful interdisciplinary relationships rather than simplistic one-to-one correspondences.

**Electronic Engineering Analysis**

Electronic Engineering showed a similar multi-pathway pattern, with General Engineering items providing the primary contribution (combined Pratt's measure = 0.445) but Computer Science items also showing notable importance (Pratt's measure = 0.118). This relationship reflects the increasing convergence of electronic engineering and computer science in modern technology applications, providing evidence that the neural network appropriately recognized these domain connections.

#### 4.2.4 Concerning Cases: Weak Theoretical Alignment

**Gender Studies Analysis**

Gender Studies represented one of the more concerning cases for MNN-1 validity, despite achieving high accuracy (0.95). The three Gender Studies items showed only moderate Pratt's measures (0.087, 0.092, 0.089), collectively explaining just 26.8% of the variance in Gender Studies accuracy prediction. The remaining variance was distributed across numerous variables without clear theoretical justification, including items from unrelated fields such as Engineering and Physical Sciences.

This pattern suggests that while MNN-1 achieved high accuracy in predicting Gender Studies preferences, this accuracy may have been achieved through spurious correlations rather than theoretically meaningful relationships. The diffuse pattern of variable importance raises concerns about the generalizability and interpretability of the network's decisions for this particular major.

**Statistics Analysis**

Statistics showed a similarly concerning pattern, with the three Mathematics items (which serve as proxies for Statistics in the CMPA short form) achieving only modest Pratt's measures (0.094, 0.101, 0.088). The network appeared to rely heavily on items from diverse fields including Business, Psychology, and even Art, suggesting that the high accuracy (0.98) might reflect statistical artifacts rather than meaningful pattern recognition.

### 4.3 MNN-2 Results: Recall-Optimized Network Analysis

#### 4.3.1 Overall Pattern Comparison

The analysis of MNN-2, optimized for recall rather than accuracy, revealed both similarities to and important differences from MNN-1 patterns. The regression models achieved slightly lower but still substantial explanatory power, with R² values ranging from 0.38 to 0.72 (mean R² = 0.57, SD = 0.14). Interestingly, the recall optimization appeared to produce more focused variable importance patterns, with fewer majors showing the diffuse patterns observed in MNN-1.

*Figure 2. Standardized Pratt's Measures for MNN-2 (Recall-Optimized Network). This heatmap displays the variable importance patterns for the recall-optimized network. Compared to Figure 1, the recall-optimized network shows more concentrated patterns of importance, with stronger diagonal relationships indicating enhanced theoretical alignment for many majors.*

#### 4.3.2 Enhanced Theoretical Alignment Cases

**Psychology Major Under Recall Optimization**

Under recall optimization, Psychology showed even stronger theoretical alignment than in the accuracy-optimized network. The three Psychology items achieved Pratt's measures of 0.225, 0.198, and 0.187, collectively explaining 61.0% of the variance in Psychology recall prediction. This enhanced focus suggests that optimizing for recall may naturally lead to more theoretically interpretable decision-making processes.

**Education Major Analysis**

Education demonstrated particularly strong theoretical alignment under recall optimization, with the three Education items achieving exceptionally high Pratt's measures (0.267, 0.234, 0.221) that explained 72.2% of the total variance. This pattern indicates that MNN-2 developed a highly focused and theoretically appropriate approach to identifying students who might be interested in Education, relying primarily on education-specific content as expected.

*Figure 3. Detailed Analysis of Standardized Pratt's Measures for Selected Majors. This bar chart visualization shows the distribution of variable importance across different predictors for key majors including Accounting, Civil Engineering, Criminology, Education, Electronic Engineering, General Business, Gender Studies, and Psychology. Each panel demonstrates the specific pattern of predictor contributions, with theoretically expected predictors showing the highest bars for most majors.*

**4.3.3 Improved Complex Pattern Recognition**

**Engineering Majors Under Recall Optimization**

The various engineering specializations showed more coherent patterns under recall optimization compared to accuracy optimization. Mechanical Engineering, for instance, demonstrated strong primary reliance on General Engineering items (combined Pratt's measure = 0.521) with meaningful secondary contributions from Mathematics (0.098) and Physics (0.087), reflecting the quantitative and scientific foundations of the field.

Chemical Engineering showed similar improvement, with General Engineering items providing the primary signal (0.487) and Chemistry items contributing meaningfully (0.142), accurately reflecting the interdisciplinary nature of chemical engineering education and practice.

**4.4 Cross-Network Comparative Analysis**

**4.4.1 Consistency Across Optimization Objectives**

Comparing results across MNN-1 and MNN-2 revealed remarkable consistency for majors with strong theoretical alignment. Psychology, Criminology, and most Education majors showed similar patterns of variable importance regardless of whether the network was optimized for accuracy or recall, suggesting that these represent robust, theoretically grounded relationships that the networks consistently discovered.

**4.4.2 Divergence Patterns and Their Implications**

However, significant divergences emerged for majors that showed weak theoretical alignment in MNN-1. Gender Studies, for example, showed markedly different patterns between the two networks, with MNN-2 demonstrating somewhat stronger (though still suboptimal) reliance on Gender Studies-specific items. This divergence suggests that the poor patterns observed in MNN-1 may indeed reflect spurious correlations rather than fundamental limitations of the neural network approach.

**4.5 Methodological Validation Results**

**4.5.1 Robustness Checks**

Sensitivity analyses using different perturbation proportions (25% and 75% variable inclusion) confirmed the stability of the main findings. The relative rankings of variable importance remained highly consistent (Spearman's ρ > 0.85) across different perturbation strategies, providing confidence in the method's reliability.

Cross-validation analyses using subsets of perturbation trials showed that Pratt's measures stabilized after approximately 3,000 trials, indicating that the full 5,000-trial protocol provided more than adequate statistical power for reliable estimation.

**4.5.2 Convergent Validity Evidence**

To validate the proposed method's results, we compared the Pratt's measure findings with alternative explanation approaches including permutation importance and SHAP (SHapley Additive exPlanations) values computed on a subset of cases. The correlation between Pratt's measures and these alternative methods was substantial (r = 0.73 with permutation importance, r = 0.68 with SHAP values), providing convergent validity evidence for the proposed approach while demonstrating its unique contributions.

**4.6 Practical Implications for Neural Network Validation**

The comprehensive results demonstrate that the proposed perturbation-Pratt's measure method successfully differentiates between valid and potentially spurious neural network behavior patterns. The method identified 38 majors where neural network decisions appeared to be based on theoretically appropriate information, 8 majors with acceptable but suboptimal patterns, and 4 majors where decisions appeared to rely on potentially spurious correlations despite high overall accuracy.

These findings have important implications for the practical deployment of neural networks in educational assessment contexts. The results suggest that high accuracy alone is insufficient for validating neural network behavior; systematic analysis of variable importance patterns is essential for ensuring that networks make decisions for the right reasons rather than achieving accuracy through statistical artifacts.

**References**

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

iKoda. (2017). *College Major Preference Assessment*. iKoda Research.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics* (pp. 245-260). University of Tampere.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275.

Wu, A. D., Hu, S. F., & Stone, C. A. (2022). Neural networks as flexible scoring mechanisms for short test forms. *Educational and Psychological Measurement*, 82(4), 687-712.

**5. Discussion**

**5.1 Principal Findings and Their Significance**

This study proposed and validated a novel explainable AI method that combines perturbation techniques with Pratt's measures to provide statistically grounded insights into neural network behavior. The comprehensive evaluation using the College Major Preference Assessment (CMPA) with 9,442 participants across 50 college majors demonstrates the method's effectiveness in differentiating between theoretically valid and potentially spurious neural network decision-making patterns.

The principal finding that emerges from this research is that high predictive accuracy alone is insufficient for validating neural network behavior in educational and psychological assessment contexts. The proposed method successfully identified instances where networks achieved impressive accuracy (e.g., 95% for Gender Studies, 98% for Statistics) while relying on theoretically inappropriate or diffuse patterns of variable importance. This discovery has profound implications for the responsible deployment of neural networks in high-stakes educational contexts where understanding the reasoning behind predictions is essential for validity and interpretability.

The method's ability to distinguish between valid and spurious patterns represents a significant advance in explainable AI for educational applications. Unlike existing explanation methods that often provide technical insights primarily useful to machine learning experts, the proposed approach generates results that are directly interpretable by domain experts in education and psychology who may lack deep technical expertise in neural network architectures.

**5.2 Methodological Contributions to Explainable AI**

**5.2.1 Integration of Classical Statistics with Modern AI**

The proposed method represents a novel bridge between classical statistical techniques and modern explainable AI approaches. By leveraging Pratt's measures—a well-established tool in regression analysis—the method provides results that connect to familiar statistical concepts while addressing the interpretability challenges of neural networks. This integration offers several advantages over purely machine learning-based explanation methods.

First, the statistical foundation provides theoretical guarantees about the interpretation of results. Pratt's measures have well-understood mathematical properties, including the crucial additive property that ensures all measures sum to the total explained variance. This characteristic enables practitioners to assess not only which variables are important but also how much of the total predictive power can be attributed to theoretically expected variables versus potentially spurious correlations.

Second, the method produces results that are immediately interpretable by researchers trained in traditional statistical methods, reducing the barrier to adoption in fields where statistical literacy is more common than machine learning expertise. The proportion of variance explained by each variable provides an intuitive metric that domain experts can readily understand and evaluate against their theoretical expectations.

**5.2.2 Advantages Over Existing Perturbation Methods**

The proposed approach differs fundamentally from typical perturbation methods in ways that address several limitations of existing techniques. Traditional perturbation methods manipulate the values of input variables, often replacing them with random values, mean values, or values sampled from other instances. This approach can produce unrealistic or impossible input combinations that may not reflect how the model would behave with naturally occurring data.

In contrast, our method disables entire variables rather than manipulating their values, ensuring that all analysis is conducted with realistic data patterns. When variables are included in the analysis, they retain their original values from real participants, maintaining the natural covariance structure and realistic value ranges that characterize the domain. This approach provides more ecologically valid insights into how the neural network actually processes real-world data.

Furthermore, the perturbation-retraining approach captures the neural network's learning process rather than just its final state. By training new networks with different variable combinations, the method reveals which variables are truly necessary for the network to discover meaningful patterns, rather than simply which variables the pre-trained network appears to weight heavily in its final form.

**5.2.3 Model-Agnostic Applicability**

A significant advantage of the proposed method is its model-agnostic nature. The approach requires no knowledge of internal network architecture, no access to gradients or weights, and no modifications to the training process. This characteristic makes the method broadly applicable across different neural network architectures and easily adaptable to emerging modeling approaches.

The model-agnostic property is particularly valuable in practical applications where practitioners may need to explain models developed by others or where the internal architecture of deployed models may be proprietary or complex. The method's ability to provide meaningful insights without requiring specialized knowledge of the underlying model makes it accessible to a broader range of users and applications.

**5.3 Theoretical Implications for Educational Assessment**

**5.3.1 Validity Evidence in Neural Network-Based Assessment**

The findings contribute significantly to the ongoing discussion about validity evidence in educational and psychological assessment, particularly as these fields increasingly incorporate machine learning approaches. Traditional psychometric theory emphasizes the importance of construct validity—ensuring that assessments measure what they purport to measure—but provides limited guidance for evaluating this concept in the context of neural networks.

The proposed method offers a concrete approach for gathering construct validity evidence for neural network-based assessments. By examining whether networks rely on theoretically appropriate input variables, practitioners can assess whether the networks are likely measuring the intended constructs or are instead capitalizing on irrelevant correlations in the data. This capability addresses a critical gap in the literature on machine learning applications in educational measurement.

The distinction between accuracy-based and validity-based evaluation represents a fundamental shift in how neural networks should be evaluated for educational applications. While predictive accuracy remains important, the current research demonstrates that accuracy alone can be misleading when networks achieve high performance through theoretically inappropriate means. The proposed method provides tools for conducting the more nuanced evaluation required for responsible deployment in educational contexts.

**5.3.2 Implications for Automated Scoring and Diagnostic Systems**

The research has particular relevance for the growing field of automated scoring and diagnostic systems in education. As neural networks are increasingly deployed for tasks such as essay scoring, automated feedback generation, and learning analytics, understanding whether these systems make decisions based on appropriate criteria becomes paramount.

The method's ability to identify cases where networks achieve high accuracy through potentially spurious means has important implications for fairness and bias in educational AI systems. Networks that rely on unexpected or theoretically inappropriate variables may inadvertently perpetuate existing biases or create new forms of unfairness that are difficult to detect through traditional validation approaches.

**5.4 Practical Applications and Generalizability**

**5.4.1 Beyond Educational Assessment**

While demonstrated in the context of college major preference assessment, the proposed method has broad applicability across domains where neural networks are used for consequential decision-making. The core principle—validating that networks make decisions based on theoretically appropriate information—is relevant to any application where domain expertise provides clear expectations about which input variables should be most important.

In medical diagnosis, for example, the method could help validate whether diagnostic neural networks rely on clinically relevant symptoms and test results rather than spurious correlations in patient data. In hiring and personnel selection, the approach could identify whether recruitment algorithms base decisions on job-relevant qualifications or potentially discriminatory factors. In financial services, the method could assess whether credit scoring algorithms rely on legitimate financial indicators rather than proxies for protected characteristics.

**5.4.2 Implementation Considerations for Practitioners**

The practical implementation of the proposed method requires careful consideration of several factors. The computational intensity of training thousands of neural networks represents a significant resource requirement that may limit adoption in some contexts. However, the computational cost should be weighed against the potential consequences of deploying inadequately validated neural networks in high-stakes applications.

For organizations considering implementation, several strategies can help manage computational requirements. First, the method can be applied selectively to the most critical outputs or decisions rather than comprehensively analyzing all aspects of a complex system. Second, the number of perturbation trials can be adjusted based on the stability of results, with fewer trials potentially sufficient for initial validation and more extensive analysis reserved for final validation. Third, the method can be implemented as part of a broader validation framework that includes other, less computationally intensive checks.

**5.5 Limitations and Considerations**

**5.5.1 Computational Requirements and Scalability**

The most significant limitation of the proposed method is its computational intensity. Training 5,000 neural networks for each target network requires substantial computing resources and time, potentially limiting the method's adoption in resource-constrained environments. The current implementation required approximately three weeks of computation time using specialized hardware, which may not be feasible for all applications.

However, this limitation should be considered in context. The computational cost represents a one-time investment for validating a neural network that may be used to make thousands or millions of consequential decisions. For high-stakes applications where incorrect decisions could have significant negative impacts, the validation cost may be justified by the insights gained and the potential problems avoided.

Future research should investigate methods for reducing computational requirements while maintaining the quality of insights. Possible approaches include more efficient sampling strategies for variable selection, early stopping criteria based on convergence of Pratt's measures, and parallel computing implementations that could reduce wall-clock time.

**5.5.2 Dependence on Theoretical Frameworks**

The method's effectiveness depends critically on the availability of clear theoretical expectations about which input variables should be most important for specific outcomes. In domains where theory is poorly developed or where multiple competing theoretical frameworks exist, the interpretation of results becomes more challenging.

This limitation is particularly relevant for exploratory applications where researchers may hope to use the method to discover unexpected relationships or generate new theoretical insights. While the method can identify which variables are most important for network performance, it cannot independently determine whether these patterns are theoretically meaningful without external domain expertise.

**5.5.3 Assumptions About Linear Relationships**

The use of linear regression and Pratt's measures in the final analysis step assumes that the relationship between variable inclusion and network performance can be adequately captured by linear models. While our results suggest that this assumption is reasonable for the current application, it may not hold in all contexts, particularly those involving complex interaction effects or non-linear relationships between variable availability and model performance.

Future extensions of the method could explore non-linear modeling approaches for the final analysis step, potentially using techniques such as random forests or neural networks to model the relationship between variable inclusion and performance. However, such extensions would need to maintain the interpretability advantages of the current approach.

**5.6 Future Research Directions**

**5.6.1 Methodological Extensions and Improvements**

Several methodological extensions could enhance the proposed approach. First, investigating optimal perturbation strategies beyond random variable selection could potentially improve the efficiency and informativeness of the analysis. Systematic exploration of different inclusion proportions, stratified sampling based on theoretical categories, or adaptive sampling strategies that focus on the most informative variable combinations could yield insights with fewer computational resources.

Second, extending the method to handle different types of neural network architectures, including convolutional neural networks, recurrent neural networks, and transformer-based models, would broaden its applicability. Each architecture type may require adaptations to the perturbation strategy or analysis approach to ensure meaningful results.

Third, developing methods for handling temporal or sequential data could extend the approach to applications such as student learning analytics, where the timing and sequence of information may be as important as the information content itself.

**5.6.2 Integration with Other Explainable AI Techniques**

Future research should explore how the proposed method can be integrated with other explainable AI techniques to provide more comprehensive insights into neural network behavior. For example, combining perturbation-Pratt's measure analysis with attention visualization, gradient-based explanation methods, or counterfactual explanation techniques could provide multiple perspectives on network behavior that triangulate on a more complete understanding.

Such integration could help address some limitations of individual explanation methods. For instance, gradient-based methods might provide insights into local decision-making patterns that complement the global variable importance patterns revealed by the proposed method.

**5.6.3 Applications to Emerging Educational Technologies**

The rapid development of AI-powered educational technologies presents numerous opportunities for applying and extending the proposed method. Intelligent tutoring systems, automated essay scoring platforms, adaptive learning systems, and educational recommendation engines all rely on neural networks to make consequential decisions about student learning experiences.

Future research should investigate how the method can be adapted to provide ongoing validation of deployed educational AI systems, potentially enabling real-time monitoring of whether systems continue to make decisions based on appropriate criteria as they encounter new student populations or educational contexts.

**5.7 Implications for Policy and Practice**

**5.7.1 Regulatory and Ethical Considerations**

The research has important implications for the regulation and governance of AI systems in educational contexts. As regulatory frameworks for AI in education continue to develop, the proposed method could contribute to technical standards for validating neural network behavior. The ability to provide concrete evidence about whether systems make decisions based on appropriate criteria could support compliance with emerging regulations requiring algorithmic transparency and accountability.

Educational institutions and technology providers may need to incorporate similar validation approaches into their development and deployment processes to ensure responsible AI use. The method provides a concrete tool for demonstrating that neural network-based educational systems operate in ways that align with educational theory and best practices.

**5.7.2 Professional Development and Training Implications**

The successful implementation of methods like the one proposed requires educational professionals who understand both domain expertise and basic principles of machine learning validation. This highlights the need for professional development programs that help educators and educational researchers develop sufficient AI literacy to effectively oversee and validate AI systems in their domains.

Similarly, machine learning practitioners working in educational contexts need sufficient understanding of educational theory and measurement principles to design and validate systems appropriately. The proposed method could serve as a bridge between these communities, providing a technically rigorous but conceptually accessible approach to validation that both groups can understand and use effectively.

**6. Conclusion**

**6.1 Summary of Contributions**

This research addresses one of the most pressing challenges in the application of artificial intelligence to educational and psychological contexts: ensuring that neural networks make decisions for theoretically appropriate reasons rather than achieving high accuracy through spurious correlations. Through the development and validation of a novel perturbation-Pratt's measure method, we have demonstrated a practical approach for explaining and validating neural network behavior that bridges the gap between machine learning performance and domain expertise.

The key contributions of this work can be summarized as follows:

**Methodological Innovation**: The integration of perturbation techniques with Pratt's measures represents a significant methodological advancement in explainable AI. By combining the flexibility of intervention-based explanation methods with the statistical rigor of classical variable importance measures, the proposed approach provides interpretable, quantitative insights into neural network decision-making processes. This methodology offers a unique solution to the longstanding tension between predictive performance and interpretability in machine learning applications.

**Empirical Validation**: The comprehensive evaluation using real-world data from 9,442 participants across 50 college majors provides robust evidence for the method's effectiveness. The analysis successfully differentiated between neural networks that relied on theoretically appropriate variables (such as Psychology items predicting Psychology major preference) and those that achieved high accuracy through potentially spurious means (such as diffuse patterns for Gender Studies prediction). This empirical validation demonstrates the method's practical utility for real-world applications.

**Theoretical Framework**: The research contributes to the theoretical understanding of validity in neural network-based educational assessment. By providing a concrete framework for evaluating whether neural networks make decisions based on construct-relevant information, the study extends traditional psychometric concepts to modern machine learning contexts. This theoretical contribution has implications beyond the immediate application domain, offering insights relevant to any field where understanding the reasoning behind AI decisions is crucial.

**Practical Applicability**: The model-agnostic nature of the proposed method makes it broadly applicable across different neural network architectures and application domains. The approach requires no specialized knowledge of internal network parameters or architecture-specific features, making it accessible to domain experts who may lack deep technical expertise in machine learning. This accessibility is crucial for widespread adoption in educational and psychological research communities.

**6.2 Implications for Practice**

The findings have immediate practical implications for researchers, practitioners, and policymakers involved in the development and deployment of neural network-based systems in educational contexts. Most fundamentally, the research demonstrates that predictive accuracy alone is insufficient for validating AI systems in high-stakes applications. Organizations considering the deployment of neural networks for educational assessment, student placement, or career guidance must implement validation procedures that examine not just whether the systems are accurate, but whether they are accurate for the right reasons.

The proposed method provides a concrete tool for conducting such validation. Educational technology companies can use the approach to verify that their AI systems rely on educationally relevant factors rather than spurious correlations that might lead to unfair or inappropriate decisions. Educational institutions can employ the method to evaluate AI systems before adoption, ensuring that automated decision-making tools align with educational theory and best practices.

For researchers developing neural network applications in psychology and education, the method offers a systematic approach for demonstrating the construct validity of their models. This capability is particularly valuable for publication in venues that require evidence of theoretical grounding and interpretability, not just predictive performance.

**6.3 Broader Impact on Explainable AI**

Beyond its immediate applications in educational assessment, this research contributes to the broader explainable AI movement in several important ways. First, it demonstrates the value of integrating classical statistical techniques with modern explanation methods, showing how established tools from statistics can enhance the interpretability of contemporary machine learning approaches.

Second, the research highlights the importance of domain expertise in interpreting explanation results. While many explainable AI methods focus on generating explanations that are technically sophisticated, the current work emphasizes the need for explanations that are meaningful within specific application domains. The proposed method's ability to connect neural network behavior to theoretical expectations provides a model for developing domain-relevant explanation approaches in other fields.

Third, the work contributes to ongoing discussions about the relationship between accuracy and interpretability in machine learning. Rather than viewing these as necessarily competing objectives, the research suggests that proper validation should examine both predictive performance and the appropriateness of the decision-making process. This perspective has implications for the development of responsible AI systems across all application domains.

**6.4 Limitations and Future Directions**

While the proposed method represents a significant advance in explainable AI for educational applications, several limitations suggest directions for future research. The computational intensity of the approach, requiring the training of thousands of neural networks, may limit its adoption in resource-constrained environments. Future work should investigate more efficient implementation strategies, including parallel computing approaches and early stopping criteria that could reduce computational requirements while maintaining the quality of insights.

The method's reliance on linear regression for the final analysis step assumes that relationships between variable inclusion and performance can be captured by linear models. While this assumption appears reasonable for the current application, future research should explore non-linear alternatives that might capture more complex relationships while maintaining interpretability.

The approach also depends on the availability of clear theoretical expectations about variable importance, limiting its applicability in domains where theory is poorly developed. Future extensions could investigate how the method might be adapted for exploratory applications where the goal is to discover rather than validate theoretical relationships.

**6.5 Vision for Future Applications**

Looking forward, the proposed method and its underlying principles could be extended to address emerging challenges in educational AI. As artificial intelligence becomes increasingly sophisticated and ubiquitous in educational contexts, the need for robust validation methods will only grow. The approach developed here provides a foundation for more comprehensive validation frameworks that could be adapted to different types of AI systems and educational applications.

Potential future applications include the validation of large language models used for automated essay scoring, the explanation of recommendation systems in adaptive learning platforms, and the interpretation of computer vision systems used for automated assessment of student work. Each of these applications would require adaptations of the basic methodology, but the core principle of validating that AI systems make decisions based on appropriate criteria remains universally relevant.

The method could also contribute to the development of AI systems that are interpretable by design rather than requiring post-hoc explanation. By incorporating validation criteria similar to those used in the proposed method into the training process itself, future research might develop neural networks that inherently make decisions based on theoretically appropriate factors.

**6.6 Final Thoughts**

The integration of artificial intelligence into educational and psychological practice represents both tremendous opportunity and significant responsibility. AI systems have the potential to provide personalized, adaptive, and highly effective educational experiences that could transform learning outcomes for students worldwide. However, realizing this potential requires more than just achieving high predictive accuracy; it demands ensuring that AI systems operate in ways that are transparent, fair, and aligned with educational goals and values.

This research contributes to that broader goal by providing tools and frameworks for understanding and validating AI behavior in educational contexts. The proposed perturbation-Pratt's measure method represents one step toward the responsible development and deployment of AI in education, offering a concrete approach for ensuring that neural networks make decisions for the right reasons.

As the field continues to evolve, the principles demonstrated here—the importance of theoretical grounding, the value of domain expertise in interpretation, and the need for comprehensive validation beyond predictive accuracy—will remain relevant regardless of the specific technologies employed. The ultimate goal is not just to build AI systems that are accurate, but to build systems that are worthy of the trust placed in them by educators, students, and society as a whole.

The path toward this goal requires continued collaboration between machine learning researchers, educational practitioners, and domain experts who can ensure that technical advances serve educational purposes. This research represents one contribution to that collaborative effort, providing both technical tools and conceptual frameworks that can support the responsible development of AI in education. As we continue to navigate the integration of artificial intelligence into educational practice, methods like the one proposed here will play a crucial role in ensuring that technological capability is matched by theoretical understanding and practical wisdom.

**References**

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 9, 2018.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

iKoda. (2017). *College Major Preference Assessment*. iKoda Research.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1-15.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international conference in statistics* (pp. 245-260). University of Tampere.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer.

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, 45(1-3), 253-275.

Wu, A. D. (2021). Validation evidence for the College Major Preference Assessment. *Journal of Career Assessment*, 29(3), 456-478.

Wu, A. D., Hu, S. F., & Stone, C. A. (2022). Neural networks as flexible scoring mechanisms for short test forms. *Educational and Psychological Measurement*, 82(4), 687-712.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818-833.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*, 559(7714), 324-326.