# Human-AI Co-Construction of Interpretable Predictive Models: The Case of Scoring Systems

Stefan Heid[1†], Jaroslaw Kornowicz[2†], Jonas Hanselle[13], Eyke Hüllermeier[13], Kirsten Thommes[2]

[1]LMU Munich
{stefan.heid,jonas.hanselle,eyke}@lmu.de

[2]Paderborn University
{jaroslaw.kornowicz,kirsten.thommes}@upb.de

[3]MCML, Munich

[†] equal contribution

This study explores the co-construction of probabilistic scoring systems. Using a self-developed web-based tool, called PSLVIS, participants were able to create their own decision-support models through an interactive interface. Seven academic advising experts participated, assessing the probability of student success both with and without the assistance of a Probabilistic Scoring List (PSL). The results indicate that while the co-constructed models slightly improved the experts' accuracy, they also increased decision time. Experts interacted with PSLVIS and PSL in diverse ways, displaying different levels of algorithmic aversion and appreciation. This study underscores the potential of decision-support systems that integrate data-driven algorithms with human expertise, while also revealing the wide range of challenges that need to be addressed for successful co-construction and practical implementation.

We would like to sincerely thank the participants of the study. We gratefully acknowledge funding by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG): TRR 318/1 2021 – 438445824.

# 1 Introduction

With the increasing access to technology and computational resources, the idea of taking advantage of Machine Learning (ML) methodology for decision support is becoming more and more feasible. Automated or partially automated decision-making with data-driven models is appealing as it can lead to more objective and accurate decisions than human decision-making alone. For example, think of decisions in the context of employee recruitment, such as hiring or placement decisions [14] in which humans alone may suffer from several biases such as "similar-to-me"-decision biases, or the data-driven construction of individualized treatment rules in personalized medicine [26].

ML models may increase the quality of decisions, but bear the problem of user acceptance: How to motivate a human decision maker to apply automated decision support systems and how to create trust and reliance in such systems [12, 13, 15]? An important prerequisite in this regard is the transparency and interpretability of the models [5,7]. Moreover, one may expect that participation, i.e., the involvement of the human expert in the process of model construction, has a positive influence, not only on acceptance [11]. Integrating humans in the process of model construction may also further improve model quality and performance — especially in cases where data is too sparse to reliably learn well-generalizing models. Hence, we introduce a *co-constructive* approach combining data-driven model induction with expert oversight.

As an underlying model class, we use so-called *scoring systems*. Roughly speaking, a scoring system proceeds from a set of (binary) features characterizing a decision context. The presence of a feature contributes a specific score (a small integer value), and a positive decision is made if the cumulative score exceeds a threshold. Models of that kind are especially comprehensible and used in many applications and fields of applied research, such as medical decision-making [18]. More specifically, we make use of PSL, an incremental and probabilistic extension of scoring systems recently developed in [10].

As a first step toward the involvement of the human expert and co-construction of a PSL, we introduce the graphical interface PSLVIS, which allows for adding, removing, and reordering features of the model as well as changing the scores.

The interface also supports the optimal (data-driven) calculation of scores and features based on the training data, thereby helping the expert to align the data with their domain knowledge. The mapping from scores to probabilities of outcomes is calculated automatically and cannot be modified. Finally, the performance of the system is visualized in the top right corner to give the user life feedback.

Building on the user interface to facilitate model co-construction, we seek to evaluate the effect of the co-constructive process on performance and reliance. More concretely, we seek to answer the following research questions:

**RQ1** How does PSL influence decision-making quality compared to humans decisions without computational support?

**RQ2** How do users interact with PSLVIS and navigate through the model space?

**RQ3** What are the thought processes and challenges users face while using PSLVIS and applying PSL?

## 2 Scoring Systems and Extensions

Scoring systems are simple linear classifiers where small integer scores are assigned to each binary feature. The sum of all scores of positive features is compared against a threshold to form a decision. PSL as introduced in [10] is an extension that produces probabilistic (instead of deterministic) predictions. Moreover, it organizes the features in the form of a decision list, so that a prediction can be made at every stage. The scores of positive features are again accumulated and then mapped to a probability estimate. An example of such a stagewise model is depicted in the bottom right of Figure 1.

Scores, feature ordering, and the probability function are learned from training data. This can be achieved by starting with an empty PSL and iteratively expanding it with the most promising feature-score-pair in a greedy fashion, similar to learning decision trees. As larger total scores should yield larger probabilities, isotonic regression is employed to obtain probability estimates

that are monotonically increasing in the total score. For a detailed description of the learning algorithm, we refer to [10].

At prediction time, features are evaluated one after another, updating the total score for each of them by adding up the scores of positive features. At each of these stages, the probability estimate can be looked up. If the estimate is not sufficiently informative to make a confident decision, additional features can be evaluated to refining the estimate and reduce uncertainty.

## 3 Co-constructive Framework: PSLvis

As a first step toward co-constructive learning of a PSL, we introduce the web interface PSLVIS instead of a purely data-driven induction. The user interface (UI) allows adding, removing, and reordering features of the model as well as changing the scores via drag-and-drop and button presses. Additionally, there are buttons to reset the model, i.e., to remove all selected features and also to add one feature optimally based on the training data. The interface also supports the optimal calculation of scores and features, allowing the experts to complement (or even replace) their expertise by a data-driven approach. The mapping from scores to probabilities is calculated automatically and cannot be modified. Finally, the performance of the entire decision list is visualized in the top right corner to give the user life feedback. A screenshot of the main view of PSLVIS is shown in Figure 1.

Significant emphasis was placed on usability during the development of the web-based UI. The UI provides an interactive experience without requiring page reloads, and any changes to features or scores result in instant model updates and performance chart adjustments. Probabilities are visually highlighted using color gradients for better clarity. The application's data model is organized into *experiments*, which can be configured independently (modifications in the user interface, different datasets, ...). Participants are assigned to these experiments, and all user data is stored in an anonymized format. All UI interactions are logged in the database, enabling a detailed analysis of the co-construction process. The implementation is publicly available[1].

[1] `https://github.com/TRR318/pslvis`

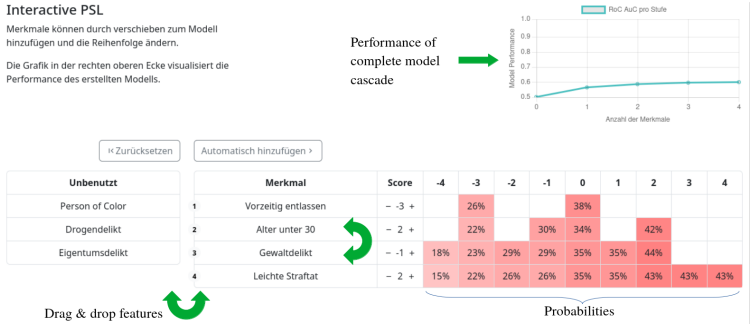


Figure 1: User interface PSLVIS, which allows adding, removing and reordering features of the PSL via drag and drop.

# 4 Method

## 4.1 Study Dataset

The study topic chosen was student counseling, specifically focusing on assessing whether a student can successfully complete their university studies. Employees from various student counseling departments were recruited as experts for the study. The basis for the study comes from the German National Educational Panel Study [2], in which pupils and students are surveyed over a longer period. This dataset is available for research purposes. We built our dataset based on Fourage and Heß [9], where we also define dropout as whether students discontinue their initial studies at their initial institution.

In our dataset, there are a total of 1,804 students, and the success rate is 65.2%. For the study, we divided the dataset according to the participants' fields of study and only used the data relevant to the areas the participants are involved with in their work. For example, participant P1 received an engineering sample, while P4 received a sample with students from law, economics, and social sciences. The dropout rate varied slightly, and the instructions within the study explained the sample.

## 4.2 Think-Aloud Method

To explore how participants interact with the co-constructive tool PSLVIS and the resulting PSL, and to identify challenges encountered during their application, we employ the think-aloud method. This qualitative research method is used to elicit cognitive processes by requiring participants to verbalize their thoughts while performing tasks, with these verbalizations recorded for subsequent analysis [4, 24]. The think-aloud method serves multiple purposes, including documenting decision-making processes [19, 23] and assessing the usability and perception of products such as software [1, 8, 22, 25]. It is also increasingly utilized in human-computer interaction research [6, 16, 20, 21].

## 4.3 Procedure

**Expert Participants.** We contacted university staff with experience in academic advising and recruited seven participants. The study took place individually and in person, with participation conducted on a computer. Experimenters were present in the room, briefly explained the procedure before the start of the study, and answered any questions for clarification. All participants signed a privacy consent form before the study began. Detailed information about the participants can be found in Table 1.

Table 1: Participant Information Table

| | **Profession** | **Major** | **Age** | **Sex** | **Sample** |
|---|---|---|---|---|---|
| P1 | Study Advisor Enginering Sci. | Education | 34 | m | Enginering |
| P2 | Study Advisor Enginering Sci. | Mech. Eng. | 34 | m | Enginering |
| P3 | Study Advisor Enginering Sci. | Ind. Eng. | 30 | m | Enginering |
| P4 | Head of Teaching/Study Center | Polical Sci. | 42 | m | Law/Eco./Social |
| P5 | Study Advisor Enginering Sci. | Mech. Eng. | 32 | m | Enginering |
| P6 | General Study Advisor | Education | 35 | f | All |
| P7 | Study Advisor Comp. Sci. | Comp. Sci. | 31 | f | Math/Nat. Sci. |

**A) Elicitation of Mental Models.** The participants' mental models regarding the decision problem are elicited. Participants rated each feature based on how they perceived the relationship between the feature and student dropout or success.

They provided a numerical rating on a scale from $-100$ (indicating dropout) to $+100$ (indicating success) to represent the perceived correlation.

**B) Probability Assessment I.** Each participant assessed the likelihood of success for 10 students. To do this, they were shown the students' features and provided a percentage-based evaluation. The 10 students were randomly selected from the eligible sample, and the order in which they were presented to each participant was randomized. No feedback was given during this stage.

**C) Co-Construction with PSLvis.** Participants then moved into a phase where they engaged with PSLVIS to co-construct PSL models. Their goal was to develop models that perform optimally within a constraint — the models could only expand up to five stages. This phase did not have a time limit, allowing participants to work through the process at their own pace. During this time, all interactions with the tool were logged, and participants were encouraged to verbalize their thought processes through the think-aloud method. Before the participants proceeded, the experimenters asked two questions: first, whether the participants were able to represent and encode their views in the model, and second, what the participants had focused on.

**D) Probability Assessment II.** In the final phase of the study, participants were asked to reassess the success probabilities of students using the PSL models they developed. This phase mirrors the initial classification task, but with the significant difference that participants could now apply their own co-constructed models. Throughout this process, the think-aloud method was employed to capture detailed insights into how participants utilize their PSL models in practice. As soon as the participants finished their second set of estimates, the experimenters asked two final questions. First, whether they had made use of the PSL levels and whether they had used all the features, and second, to what extent the PSL had influenced their decisions.

# 5 Results

## 5.1 Participants' Assessments

Table 2 presents the average times all participants took to make their assessments and their accuracy, measured by the Brier score (lower is better) [3]. The results are divided between the two assessment rounds. A purely data-driven PSL model, evaluated using individual samples for each participant, serves as the reference for accuracy.

Although a precise statistical evaluation is not possible due to the small sample size, the descriptive analysis shows that experts took longer to make their assessments in the second round. This is likely because they were also interested in reviewing their own PSL models, though there is considerable variance in this aspect. In terms of accuracy, experts generally performed slightly better with the PSL model than without, though this was not true for everyone. The reference values indicate that, on average, the experts outperformed the purely data-driven model in the second round.

Table 2: Average duration for the assessment of the students in seconds and the Brier scores (lower is better) for the first and second assessments. The PSL column serves as a reference for a purely data-driven model. The bottom row shows the average for all.

| | Average Time | | Brier Score | | |
|---|---|---|---|---|---|
| | **I** | **II** | **I** | **II** | **PSL** |
| P1 | 110.8 | 70.8 | 0.29 | 0.28 | 0.26 |
| P2 | 62.6 | 58.6 | 0.32 | 0.23 | 0.26 |
| P3 | 75.4 | 94.1 | 0.24 | 0.24 | 0.26 |
| P4 | 54.3 | 75.4 | 0.29 | 0.24 | 0.27 |
| P5 | 45.2 | 31.6 | 0.33 | 0.26 | 0.26 |
| P6 | 42.1 | 36.9 | 0.14 | 0.20 | 0.29 |
| P7 | 19.2 | 104.9 | 0.26 | 0.25 | 0.25 |
| **ø** | **58.51** | **67.46** | **0.27** | **0.24** | **0.26** |

## 5.2 Co-construction as Navigation in the Model Space

In phase A) of Section 4.3 the participants were asked to express their mental model by providing weights for each feature in the dataset to elicit positive or negative correlation with the target class "study success". Figure 2 shows the features and the accompanying assigned scores. The features are sorted by the mean absolute score of the participant's mental model assessments, shown as blue bars. The participants assume that neuroticism is the strongest indicator for study dropout, while life satisfaction, consciousness, and openness are the three strongest indicators for study success.

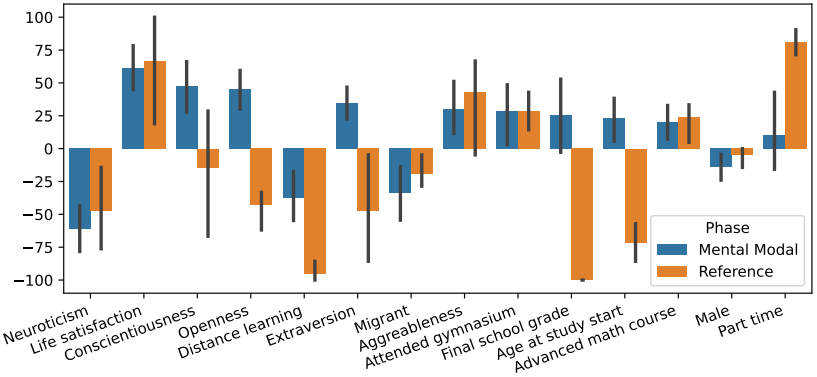


Figure 2: The blue bars show the mean feature importance assessment from phase A of the study; the orange bars show the mean score for that features when PSL is fitted on the respective data sample, normalized to the same domain $[-100, 100]$. The error bars show the 95% confidence interval of the mean.

The orange bars show the average scores of a fully data-driven PSL trained, each on the same dataset as the participant. For easier comparability, the scores from $\{-3, \ldots, +3\}$ have been rescaled to $[-100, 100]$. All dataset samples have been pooled in that figure, as the number of participants is so small. The participant's assessment of feature importance strongly disagrees with the purely data-driven feature importance as calculated from the PSL scores. In the reference model, poor final school grades, distance learning, and high age at the start of study are the strongest indicators of study dropout, while studying part-time and having high life satisfaction are the strongest indicators of success. Since the

participant's goal was to have a high predictive performance on data points from this dataset, it is important to lower the model gap between the mental model and the data distribution in the domain.

During the co-construction process, features and scores can be changed. Each of these changes can be interpreted as an action that navigates from one model $h$ to another model $h'$ with edited features and scores. Hence, the co-constructive process can be seen as a navigation in the space of PSL models. We define the following distance function between PSL models $h$ and $h'$ in order to analyze how human co-constructors navigate through this space as follows:

$$d(h,h') = Kendall\big(F(h),F(h')\big) + \left\| \frac{S(h)-S(h')}{|\mathcal{S}|} \right\|,$$

which is the sum of the Kendall $\tau$ distance of the feature rankings and the $L_2$-norm of the normalized score difference ($\mathcal{S}$ is the set of possible scores). $F(h)$ denotes the feature ranking[2] and $S(h)$ the score assignments[3] of $h$ .

**Model changes during co-construction.** The model changes during co-construction can be analyzed by comparing the current model, created by the participant, to other models. To this end, the distance between the mental model and the purely data-driven model was observed. As the mental model of part A) of the study is only observed through the feature importance scores from $[-100,+100]$, a PSL can be constructed as follows: First, the features are sorted with regard to the absolute importance score in descending order. Ties of feature importance assessments are broken arbitrarily. Second, the scores can be computed by mapping the $[-100,+100]$ interval to the score set $\{-3,\ldots,+3\}$ by rescaling linearly and rounding.

Figure 3 shows the relative distance of the co-constructed model towards the mental model and the data-driven reference model over the time of the co-constructive process. All participants except P1 and P6 have an overall trend towards the data-driven model, starting with a model that is closer to their initial belief. For participants P2, P3, and P5, the final model is especially close to the data-driven model at the end of the co-construction phase. The large steps

[2] Features not present in $h$ are assigned the maximum rank $|\mathcal{F}|$, with $\mathcal{F}$ being the set of all features.
[3] The score of absent features is set to 0.

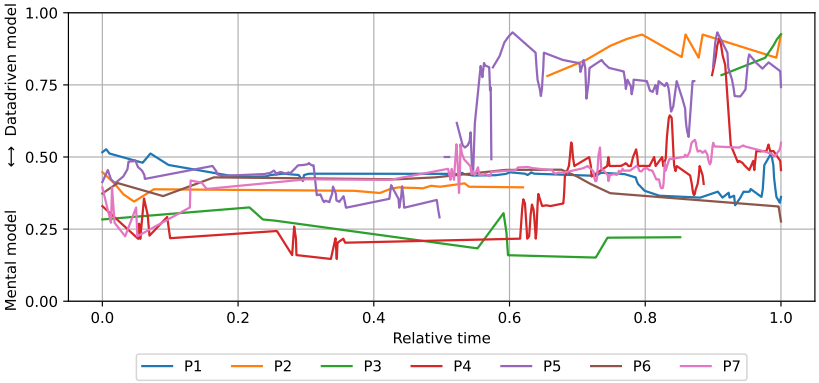


Figure 3: The relative distance between the co-constructed model for each participant towards two reference models is shown on the y-axis: one model created from the feature importance assessment ($y = 0$) and one model trained purely data-driven ($y = 1$). The x-axis shows the relative time over the course of the co-constructive process.

towards the data-driven model in P2 through P5 are caused by the participants' use of the reset and automatic feature addition buttons. However, P7 also co-constructed the model closer to the data-driven model only by manually adding features and modifying scores. When ignoring changes induced by the automatic feature addition, we can see that most participants seem to end up with models that have similar distances to their initial mental model and the data-driven reference. This is particularly illustrated with P4, where the changes from the automatic feature addition after around 90% of the co-construction time are mostly reverted manually. Similarly, P5 also modifies the model after feature addition to move closer toward their mental modal after using automatic feature addition (60%, 90% time). As Figure 3 only visualizes the relative distance to two anchor points, it still seems that most participants do not fully explore the space of models, as the relative distance changes are relatively small. Note that all co-constructive models consist of at most 5 features, while the two reference models consist of all features.

## 5.3 PSLvis User Actions

Figure 4 illustrates how the participants interacted with PSLVIS during the co-construction process. This is shown through a timeline for each participant, revealing several key insights: the duration of the co-construction process varied significantly. While two experts (P2, P6) spent less than 5 minutes on this part, two others (P4, P7) took more than 13 minutes.


Figure 4: Action timelines for each participant, showing the time elapsed since the first recorded action. Each marker represents the subject's specific action.

All participants started by independently adding features to the model. Three of them simultaneously adjusted the scores (P3, P4, P7), while the others first focused on filling in the model. Participant P6 did not remove any features and stayed with the initially selected ones. Four participants (P2, P3, P4, P5) used the reset function, all directly related to the automatic addition of features. Of these, one participant (P3) accepted the features without adjusting the scores, two (P2, P4) only modified the scores, and one (P5) both changed the features and adjusted the scores. One expert (P7) used the automatic addition function without resetting.

## 5.4 Think-Aloud Results

The audio data was first transcribed and then inductively coded after multiple readings. First, the data was categorized into statements about PSLVIS and/or PSL, and second, into statements about thought processes and/or challenges.

## 5.5 Co-Construction with PLSvis

**Thought Processes.** Participants in the co-construction process with PSLVIS engage in various strategies as they explore and modify the model. They add features they believe are important, sometimes based on their intuition or domain knowledge. However, they also experiment with different feature combinations and observe how these changes impact the model's performance: *"I'll throw in what I think might be important. Maybe I can also just throw in a lot and delete it afterward."* (P1). Performance is constantly evaluated, and features are removed if they do not contribute positively to the results. In some cases, participants experiment with features even if they do not fully understand them (e.g., the "Life Satisfaction" feature) just to observe how the performance changes.

Additionally, scores are tested to understand their influence on the performance: *"I can still tweak the scores a bit, but no matter what changes I make, the model performance always gets worse."* (P2)

The tool's ability to automatically suggest features is also tested, and while these suggestions may not always align with the participant's intuition, they may still be retained: *"I wanted something to be added automatically, and then it gave me 'Agreeableness'. That's a trait I haven't thought much about, but it can certainly make sense."* (P7). Throughout the process, participants remain mindful of the five-feature limit, which shapes their decisions about feature inclusion and removal: *"I would have liked to add more than five traits, but I'm not sure if that had made it more accurate."* (P7).

**Challenges.** Several challenges emerged during the co-construction process. An expert encountered features that are rare in practice, such as "Part-Time Studies While Working," which created confusion about their relevance: *"I actually noticed during the modeling process that I disagreed with at least*

*one selection of traits, because it was about a part-time study program. If I remember correctly, none of the students were actually studying part-time. That was a trait I only included because it significantly improved the model's performance. In hindsight, I think I would choose against it. This means I definitely didn't blindly follow the model, because I noticed this issue while working with it."* (P7).

Additionally, problems arose when thresholds led to scores that appeared counterintuitive, causing frustration as the participants struggled to understand why a certain threshold resulted in an "unnatural" decision boundary.

One expert expressed a desire to revise their models during the second estimation phase: *"You can't go back. Damn! I should have... Ugh, crap. I should have actually given a minus point for 'Migrant'."* (P1). There were also concerns about model performance, with some participants perceiving the performance as suboptimal. Many felt that the limit of five features was too restrictive for building effective models: *"It's incredibly difficult now with these five things I've chosen. I do believe that they are all relevant, but so is the rest. At least in part."* (P4).

For an expert, it is not clear how high or low the scores can be set (presumably due to the previous example explanation, where the scores only went up to $+2$): *"I'll play around a bit with the scores. I can do them too. I somehow thought I could only make it up to plus and $-2$, but I can make them up to seven. That's relevant, of course."* (P7).

Some experts noted discrepancies between the data provided and their real-world experiences, further diminishing confidence in the tool: *"Uh, difficult. I generally found it challenging to align my experience from my specialized counseling sessions with the traits you have. So, the selection of traits wasn't really good. I would rarely classify my counseling sessions based on what you have."* (P5). Challenges also arose with binary features; for example, when a student was female, participants found it unclear how to use the feature 'Male'. Finally, the direction of certain features, such as 'Final school grade,' created confusion, as the relationship between the feature and the score did not always align with the participant's expectations.

## 5.6 Decision-Making with PSL

**Thought Processes.** When using PSL, experts tend to go through the process methodically, often calculating probabilities all the way to the end. They adjust the output on occasion, but not always; in some cases, they accept the PSL-generated probability as is. One reason for adjusting the output was that the expert had a different weighting of features in mind compared to the system: *"Okay, I tried it with the model, and it would be 62%. When I think about it now: 18 years old, relatively young, 2.5 final grade — let's say an average school diploma. Male. Not a migrant, took advanced math courses in school. (...) Yeah, I can see again in my own evaluation that, as I said, I tend to rate all these soft skills or personality traits lower than I probably should."* (P3).

PSL influenced the estimation behavior of the participants. One expert noted that they felt motivated to deviate more from the average value when they saw the PSL probabilities, suggesting that the tool impacted their decision-making strategy: *"And if the model now gives me 86%, I'm actually more motivated, let's say, to deviate a bit more from this average score than before. So, I'll go with 75%."* (P3).

Some participants were not concerned about small differences in probabilities; minor variations did not affect their overall judgment: *"In the end, it doesn't really matter whether someone has a 75% or 85% probability of success. But it definitely makes a difference whether they have 40% or 75%."* (P4)

**Challenges.** One notable challenge with PSL was the inability to modify the model during the second estimation phase. Another challenge arose from the fact that a 0% probability is practically impossible in real-world scenarios. For one expert, receiving such a result led to significant aversion: *"The probability of successfully completing an engineering degree will never be 0%, because, well, if you have enough people, someone will always manage to do it. So, in this case, I would deviate significantly from the model and estimate it around 60%."* (P3).

# 6 Discussion

In this study, we focused on the interactive co-construction of interpretable predictive models, specifically through the lens of probabilistic scoring systems. To this end, we developed a web-based user interface that allows experts to construct their own PSL models and co-construct them with the PSL model. In a study involving 7 experts, we investigated how PSL influences the decision-making quality of users, how the experts co-construct their models, and how the interaction unfolds, identifying where challenges arise.

First, the results show that co-construction can slightly improve experts' performance in terms of accuracy, although at the cost of longer decision times. Notably, the co-constructed models also outperformed purely data-driven models. While we expected performance improvements due to co-construction and anticipated longer decision times due to the interpretability and computational complexity of PSL, the slightly better performance compared to the data-driven model can be explained by the complexity of the decision problem and the limited dataset. This also highlights that co-construction can offer an advantage, though this was not the case for all participants.

It is also important to note that there were different forms of co-construction. Some participants shifted from their own mental models towards the data-driven model, while others were resistant to the automated assistance [7]. This was evident in the think-aloud results: experts initially relied on their own opinions but experimented with different combinations of features and scores, occasionally guided by the automated function, even if they did not fully understand it. This corresponds to the issue of over-reliance or automation bias, often observed in human-AI interactions [17]. Participants partially relied on the PSL, not blindly, but taking it as advice that influenced their own judgment. However, there was aversion when the advice deviated too much or seemed unrealistic.

Our study also highlights challenges that can arise in human-AI interaction research, which may not be immediately apparent to researchers during development. For example, difficulties in understanding feature thresholds or the

binary nature of features, especially when the data in experiments does not match real-world practice.

# 7 Limitations and Future Research

A key limitation is the small number of participants, preventing statistical analysis of how PSL impacted expert decisions. This is common in human-computer interaction research with experts. Future studies might consider using laypeople via platforms like Prolific, requiring familiar datasets and problems. Although not experts, a larger sample would be more cost-effective.

Another issue is the dataset used. Estimating academic success and dropout rates is complex, and the available data was limited, resulting in low model accuracy and minimal expert improvement. Future studies could benefit from better data to enhance model performance and highlight interaction effects.

Additionally, this study did not explore how participants handle missing information during decision-making, a key focus of PSL. We kept all information available to simplify the decision problem. Future research could examine how participants manage missing data or time pressure, where they have all the information but limited time to assess everything, possibly requiring more experience with PSL.

This study highlights both the potential advantages and the challenges of co-constructed and interpretable machine learning models in decision support. While the results suggest that models created by experts can slightly improve the accuracy of their decisions, they also require significantly more time for decision-making. The co-constructive interaction with the web-based tool we developed was highly varied in terms of how the functionalities were used and experimented with, as well as in the adoption of algorithmic suggestions and the adaptation of models to individual mental models. However, some issues should be addressed in future research.

## References


[1] Obead Alhadreti and Pam Mayhew. Rethinking thinking aloud: A comparison of three think-aloud protocols. *Proceedings of Conference on Human Factors in Computing Systems*, 1-12, 2018.

[2] Hans-Peter Blossfeld, Hans-Günter Roßbach, and Jutta von Maurice. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft: Sonderheft*, 14, 2011.

[3] Glenn W. Brier. Verificiation of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1-3, 1950.

[4] Elizabeth Charters. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education Journal*, 12(2):68–82, 2003.

[5] Lingwei Cheng and Alexandra Chouldechova. Overcoming algorithm aversion: A comparison between process and outcome control. *Proceedings of Conference on Human Factors in Computing Systems*, 1-27, 2023.

[6] Michael Chromik, Malin Eiband, Felicitas Buchner, et al. I think I get your point, AI! The illusion of explanatory depth in explainable AI. *Int. Conference on Intelligent User Interfaces*, 26, 307–317, 2021.

[7] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

[8] Mingming Fan, Yiwen Wang, Yuni Xie, et al. Understanding how older adults comprehend COVID-19 interactive visualizations via think-aloud protocol. *Int. Journal of Human–Computer Interaction*, 39(8):1626–1642, 2023.

[9] Didier Fouarge and Pascal Heß. Preference-choice mismatch and university dropout. *Labour Economics*, 83, 102405, 2023.

[10] Jonas Hanselle, Johannes Fürnkranz, and Eyke Hüllermeier. Probabilistic scoring lists for interpretable machine learning. *Proceedings of DS 23rd International Conference on Discovery Science*, Springer: 189-203, 2023.

[11] Jaroslaw Kornowicz and Kirsten Thommes. Algorithm, expert, or both? Evaluating the role of feature selection methods on user preferences and reliance. *arXiv:2408.01171*, 2024.

[12] Olesja Lammert, Birte Richter, Christian Schütze, et al. Humans in XAI: increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics*, 3: 1377075, 2024.

[13] Jörg Papenkordt, Axel-Cyrille Ngonga Ngomo, and Kirsten Thommes. Are numbers or words the key to user reliance on AI? *Academy of Management Proceedings*, Vol. 2023, No. 1: 12946, 2023.

[14] D. Pessach, G. Singer, D. Avrahamia, et al. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134: 113290, 2020.

[15] Tobias M. Peters and Roel W. Visser. The importance of distrust in AI. In Luca Longo, editor, *Explainable Artificial Intelligence - First World Conference*, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part III, volume 1903 of Communications in Computer and Information Science: 301-–317. Springer, 2023.

[16] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, et al. Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. *Proceedings of Int. Conference on Intelligent User Interfaces*, 28:379–396, 2023.

[17] Max Schemmer, Niklas Kuehl, Carina Benz, et al. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of Int. Conference on Intelligent User Interfaces*, 28:410–422, 2023.

[18] Albert J. Six, Barbra E. Backus, and Johannes C. Kelder. Chest pain in the emergency room: value of the heart score. *Netherlands Heart Journal*, 16:191–196, 2008.

[19] Paul Solomon. The think aloud method: A practical guide to modelling cognitive processes. *Information Processing & Management*, 31(6):906–907, 1995.

[20] Richard Stromer, Oskar Triebe, Chad Zanocco, and Ram Rajagopal. Designing forecasting software for forecast users: Empowering non-experts to create and understand their own forecasts, *arXiv:2404.14575*, 2024.

[21] Geletaw S. Tegenaw, Demisew Amenu, Girum Ketema, et al. Evaluating a clinical decision support point of care instrument in low resource setting. *BMC Medical Informatics and Decision Making*, 23(1): 51, 2023.

[22] Thomas Van Gemert, Kasper Hornbæk, Jarrod Knibbe, and Joanna Bergström. Towards a bedder future: A study of using virtual reality while lying down. *Proceedings of Conference on Human Factors in Computing Systems*: 1–18, 2023.

[23] Jacqueline Whalley, Amber Settle, and Andrew Luxton-Reilly. A think-aloud study of novice debugging. *ACM Transactions on Computing Education*, 23(2):1–38, 2023.

[24] Michael D. Wolcott and Nikki G. Lobczowski. Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, 13(2):181–188, 2021.

[25] Xuesong Zhang and Adalberto L. Simeone. Using the think aloud protocol in an immersive virtual reality evaluation of a virtual twin. *Proceedings of Symposium on Spatial User Interaction*: 181–188, 2022.

[26] Y. Zhao, D. Zeng, A.J. Rush, and M.R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.