Doctoral thesis submitted to
the Faculty of Behavioural and Cultural Studies
Heidelberg University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy (Dr. phil.)
in Psychology

Title of the publication-based thesis
Machine Learning Applications in Psychotherapy Research

presented by
Paul Schröder-Pfeifer

year of submission
2020

Dean: Prof. Dr. rer. nat. Dirk Hagemann
Advisor: Prof. Dr. phil. Svenja Taubner

Table of contents

Research Articles for Publication

The present dissertation is based on the following research articles:

I. Evers, O., & **Schröder, P.** (2018). One Size Fits All? Using Psychosocial Risk Assessments to Predict Service Use in Early Intervention and Prevention/One size fits all? Die Eignung von Risikoscreenings zur Prognose der Inanspruchnahme von Angeboten der Frühen Hilfen. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *67*(5), 462-481. IF: 0.513

Declaration of author contributions: Evers, O.: Conceptualization, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Schröder, P.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Methodology, Software, Formal analysis

II. Georg, A.K., **Schröder-Pfeifer, P.,** Cierpka, M. & Taubner, S. (Under review). Parenting stress in the face of early regulatory disorders in infancy – what matters most? *Journal of Developmental and Behavioural Pediatrics*. IF: 2.056

Declaration of author contributions: Georg, A.K.: Conceptualization, Methodology, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Schröder-Pfeifer, P..:** Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Formal analysis, Visualization **Cierpka, M.:** Conceptualization, Supervision, Project Administration, Funding Acquisition **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

III. **Schröder-Pfeifer, P.**, Talia, A., Volkert, J., & Taubner, S. (2018) Developing an assessment of Epistemic Trust: a research protocol. *Research in Psychotherapy: Psychopathology, Process and Outcome.* IF: -

Declaration of author contributions: Schröder-Pfeifer, P.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Talia, A.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Volkert, J.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

IV. **Schröder-Pfeifer, P.**, Georg, A.K., Talia, A., Volkert, J., Ditzen, B., & Taubner, S. (Under review) The Epistemic Trust Assessment (ETA) – An experimental measure of Epistemic Trust. *Psychoanalytic Psychology.* IF: 0.958

Declaration of author contributions: Schröder-Pfeifer, P.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Georg, A.K.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Talia, A.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Volkert, J.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Ditzen, B.:** Conceptualization, Writing – Review & Editing Resources **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

V. Zettl, M., Back, S.N., Taubner, S., & **Schröder-Pfeifer, P.** (Under review) Assessing Personality Functioning and Maladaptive Traits in Young Adults: A Machine Learning Approach. *Journal of Personality Disorders.* IF: 2.440

Declaration of author contributions: Zettl, M.: Conceptualization, , Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Project Administration, Funding Acquisition **Back, S.N.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Taubner, S.:** Writing – Review & Editing, Supervision **Schröder-Pfeifer, P.:** Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – Original Draft, Writing – Review & Editing

VI. Evers, O., **Schröder-Pfeifer, P.**, Möller, H., & Taubner, S. (Under review) The Impact of Trainee Attributes and Training Variables on Competence Deterioration: Results from a Longitudinal Study in Naturalistic German Psychotherapy Training. *Training and Education in Professional Psychology*. IF: 1.028

Declaration of author contributions: Evers, O.: Conceptualization, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, **Schröder-Pfeifer, P.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Methodology, Software, Formal analysis, Visualization **Möller, H.:** Conceptualization, Funding Acquisition, Supervision **Svenja Taubner:** Conceptualization, Funding Acquisition, Project Administration, Writing – Review & Editing, Supervision

1. Introduction

Prediction of outcome or diagnoses from intake data or assessing the importance of variables as either risk factors or protective factors are fundamental tasks in psychotherapy research, in order to help clinicians and researchers to evaluate and improve treatments. With regard to data analytic assessment, these tasks can be handled by a range of parametric approaches such as regression models. However, there are cases where parametric approaches are either not applicable or have severe limitations (e.g. Strobl et al., 2009). Also, there is increasing support to the notion that biopsychosocial contributions to psychopathology are complex and cannot be sufficiently explained by a small number of variables restricted to linear relationships (Franklin, 2019; Kendler, 2019). Machine Learning (ML) algorithms offer an additional suite of methods able to deal with such complexity and can be used to extend the toolbox of psychotherapy researchers. The aim of the dissertation is to provide an understanding of machine learning application for psychotherapy research and to foster the motivation to use and improve these methods in future research.

Psychotherapy research questions often include complex relationships in high dimensional data. Highly dimensional data with interactions is generally an area where classification and regression trees (CART, Breiman, 1998), Random Forests (RF, Breiman, 2001a) and other algorithms do well, because they are able to model deep interactions between variables. Depth in this context refers to the number of involved variables, with two involved variables depth being equal to two, three involved variables being equal to depth three and so on. Within classical statistical parametric models it is usually not feasible to model interactions beyond a depth of around three, as the associated main effects and lower-level interactions (i.e. $x * y$and $y * z$ for a depth three interaction $x * y * z$) quickly lead to too many parameters for the model to converge (Strobl et al., 2009). Also, while the functional form of the

interaction pattern is usually restricted to being linear in classical approaches, algorithms can approximate almost any functional form (Strobl et al., 2009). This is often a benefit when predicting future responses, since those are often more complex than simple, linear associations. An additional benefit is that the commonly used ordinal variables in psychotherapy research do not have to be treated as being measured on an interval scale (Strobl et al., 2009).

These attributes make ML algorithms highly suited for research of risk and/or resilience factors. This is especially true in fields where there is no empirically validated theory concerning these factors, or if the theory describes a complex interaction pattern between several factors (e.g. suicidal behavior, Franklin et al., 2017; or early regulatory disorders, see Papoušek, Schieche, Wurmser, & Barth, 2004). A field that dramatically highlights the potential of ML applications is the prediction of suicide attempts. Suicide is one of the leading causes of death and a major health issue (World Health Organization, 2019), yet prediction of suicidal behavior has been a very challenging endeavor with marginal success (see Franklin et al., 2017). One of the major problems within this field of research has been that even the best isolated predictors are inaccurate (Ribeiro et al., 2016). Investigating more complex interaction patterns of potential risk factors or ranking a large number of risk factors by importance in a single study has rarely been done (Franklin et al., 2017). Because the majority of studies in the field utilized classical regression models without regularization or shrinkage, most studies were restricted to testing few predictors in isolation and did not provide variable importance statistics (Ribeiro et al., 2016). In contrast, a study by Walsh and colleagues (2017) displays the use of ML algorithms in a sample of 3250 patients with suicide attempts and 1917 patients with a history of non-suicidal self-injury, analyzing a wide range of 1328 predictor variables from electronic health records. Also, Walsh and colleagues (2017) used longitudinal data to show how the importance of different predictors shifts over time. The resulting ML model accomplished an area under the receiver operating curve (AUC) of .80 to 84. More importantly, the sensitivity of the model was as high as 96%, i.e. the algorithm correctly identified 96% of all suicidal cases ahead of time. This has proven to be superior to a classical multiple logistical regression

approach for the same data, achieving AUC values ranging from 0.66 to 0.68 (Walsh, Ribeiro, & Franklin, 2017). Other studies comparing ML models with logistic regression, aiming at predicting suicidal ideation (Ribeiro, Huang, Fox, Walsh, & Linthicum, 2019), non-suicidal self-injury (Fox et al., 2019) or differentiating between suicidal ideation and suicide attempts (Huang, Ribeiro, & Franklin, 2019), found similar results: ML methods outperformed logistic regression and the relationships between predictor variables where found to be non-linear, thus favoring ML.

Still, some studies have found no advantage of using ML models compared to classical models. A recent systematic review (Christodoulou et al., 2019) investigating uses of ML algorithms and regression models for clinical prediction modeling and assessed the risk of bias focused on methodological issues of model development, calibration, and the comparison of model performance. No benefit of ML applications was found in studies with low risk of bias (logit mean AUC difference: 0.00 [95% CI: -0.18 – 0.18]) while in studies with high risk of bias the ML models were significantly better than their regression counterparts logit mean AUC difference: 0.34 [95% CI: 0.20 – 0.47]) (Christodoulou et al., 2019). Studies with high risk of bias either used overoptimistic methods of model validation or did not report their model validation or model building in sufficient detail (Christodoulou et al., 2019). Nevertheless, it is worth mentioning that the lack of difference in the group of low bias studies can be attributed to both CART and artificial neural nets performing significantly worse than their regression equivalents (logit mean AUC difference: -0.34 [95% CI: -0.65 – -0.04] and logit mean AUC difference: -0.12 [95% CI: -0.35 – 0.12] respectively), indicating a differential effect with regard to the applied algorithm. This might also be a factor of the relative novelty of ML in psychological research, as CART are a relatively basic algorithm that is strictly inferior in almost all applications to its successor RF (see section 2), and artificial neural nets are notoriously hard to calibrate (Martinez, Black, & Romero, 2017).

Several factors can account for these different findings, among them the heterogeneous sample sizes, as well as variable quality and quantity and which algorithms and form of model validation was used. Additionally, meta-analyses and reviews (Aafjes-van Doorn, Kamsteeg, Bate, & Aafjes, 2020;

Christodoulou et al., 2019) come to the conclusion that while ML methods hold potential, reporting of methodology and findings is often lacking in critical aspects such as model validation. This can be exemplified using the abovementioned study by Walsh and colleagues (2017) where the description of the logistical regression alternative to the ML approach was lacking any model development parameters, such as which variables where included and why, making the study biased towards the ML approach.

Summarizing, while the ability of ML to model complex non-linear interactions could be a valuable asset, ML techniques are still a novelty in psychotherapy research and studies utilize ML with mixed results. Consequently, more studies have to be conducted to investigate which methodology work best in which circumstances (Aafjes-van Doorn et al., 2020; Christodoulou et al., 2019; Jacobucci, Littlefield, Millner, Kleiman, & Steinley, 2020). The aim of the present dissertation project was to apply a variety of algorithms to a wide range of clinical problems. By exploring the use of ML techniques as well as tests of generalization the author aims to contribute towards making these methods more understandable, familiar, and accessible to psychotherapy researchers. The next section will describe the rationale behind two commonly used algorithms starting with (1) classification and regression trees (CART, Breiman, 1998), including discussing its extension called Random Forest (RF, Breiman, 2001a), and (2) gradient boosting machines (GBM, Friedman, 2001). Along with examples based on the freely available data sets in the open source environment R (R Development Core Team, 2017), minimal technical explanations will be provided in addition to discussion of potential areas of application within psychotherapy research. The corresponding R code is provided in the supplements. A summary of important features and areas of application, along with potential drawbacks or pitfalls of algorithmic modeling, follows the introduction before the studies used in this dissertation project are summarized.

2. Machine learning

ML describes algorithmic statistical models, contrasting against dominant stochastic data models (Breiman, 2001b). The latter assumes that the response data is generated by a given stochastic data model, while the earlier considers the data generating mechanism to be unknown and instead tries to model the response given the inputs. There are two broad categories of ML algorithms: *Supervised* learning algorithms, where the response is known for the data and the algorithm aims at learning from the data to predict the response of new data; and *unsupervised* learning, where the response is unknown and the algorithm aims at organizing or describing the data (see Hastie, Tibshirani, & Friedman, 2009).

An important difference between ML methodology and other more commonly used statistical methods, such as linear or logistic regression, is the absence of *p*-values and in-sample mode fit as a measure of "success". Instead, with ML approaches, the main statistic of interest is the estimated prediction accuracy of the algorithm in a hold-out sample via cross-validation (CV). The accuracy for numerical outcomes is often reported as either the root mean squared error (RMSE) over all predictions against the empirical observations, or as the absolute error (AE), the absolute value of subtracting the predictions from the empirical observations. For categorical outcomes the accuracy is usually reported as the accuracy, sensitivity and specificity computed from the classification matrix of the predicted against the empirical labels of the observations, or as AUC.

It is important to notice that prediction accuracy is relative. For example, for a balanced classification task (i.e. where 50% of the sample represents the positive class, and 50% represents the negative class) in a field where no prior studies exist, an accuracy of 65% might be considered good. However, in an unbalanced classification task (i.e. where the proportion of positive or negative cases outweighs the other) such as personality disorder (PD) diagnostics where 90% of cases do not have a personality disorder, any accuracy below 90% is useless for prediction purposes. This is the case since an equivalent accuracy can be achieved without any data at all, by just classifying every new patient to be diagnosed as the more prevalent category, in this case "no PD". The rate of the more prevalent class is thus called the no information rate (NIR). However, the NIR itself does not convey enough information to

categorize a classification accuracy as good. For example, if one were to use ML to label transcripts of motivational interviewing sessions, with the main objective of finding a model which provides the most accurate labels, any accuracy lower than the 75.1% that Idalski Carcone and colleagues (2019) achieved would be worthless, even though the accuracy might lie far above the NIR. However, in a field such as suicide prevention where a correct classification might help clinicians identify patients at risk, an increase in accuracy of 1% above the NIR, even though not significant, might be valuable.

ML algorithms are usually conducted in two steps: Training the algorithm, and testing the results for generalizability. In the training phase, researchers aim at finding a good balance in calibrating their algorithm to patterns in the data specific for the groups to be analyzed to obtain accurate predictions, and not fitting too close to random noise inherent in the data, i.e. overfitting. In the test phase, the accuracy of the predictions made by the algorithm is computed by feeding the algorithm a sample different to the training set and comparing the prediction made for the new data with the actual values observed in the new sample. It is important to note that this step has to be done with a different sample (this is called the test-sample) as the one the algorithm has been trained on (thus called the training sample), as this would result in overly optimistic estimates of generalizability based on overfitting. Since many datasets used for psychotherapy research might be too small to feasibly split them into a training and test sample, k-fold CV is an alternative. In this approach the data is split into *k* folds (typically 5 or 10) and each fold is, in turn, left out of the training procedure and used to validate the results of training. The resulting accuracies are then averaged and provide a stable estimate of external generalizability (Hastie et al., 2009). A special case of k-fold validation known as leave-one-out CV is $k = N$. In this scenario, the prediction error estimate is approximately unbiased but has high variance because the different training sets a so similar (varying by only one observation). In contrast, lower numbers of *k* such as 5 or 10 have lower variance since the training sets have similar sizes as the full set, but the estimate of prediction accuracy can be biased. Generally, 5- or 10-fold CV has been shown to be a good compromise (Hastie et al., 2009).

2.1. Random Forest / Classification and Regression Trees

Random Forest (RF) developed by Breiman (2001a) is one of the older algorithms that is used in many fields and applications, such as genetic sequencing, medical and psychological diagnostics and psychotherapy research (Lee, Maenner, & Heilig, 2019; Schmitgen et al., 2019). RF are based on an even older algorithm, Classification and regression trees (CART) (Breiman, 1998). As the name implies, CART can be utilized for both regression and classification. Since regression trees and RF by extension can be explained visually, a short example utilizing the freely available "Blackmore" dataset from the R Package "carData" will be presented. The data includes 138 teenage girls, hospitalized for eating disorders and labeled "Patient", and 98 healthy control subjects. Originally, the data included multiple time points, but for simplicities sake only one measurement will be used. In this example the subjects will be classified into either patients or control predicted by their age and the amount of exercise in which the subject engaged, in hours per week.

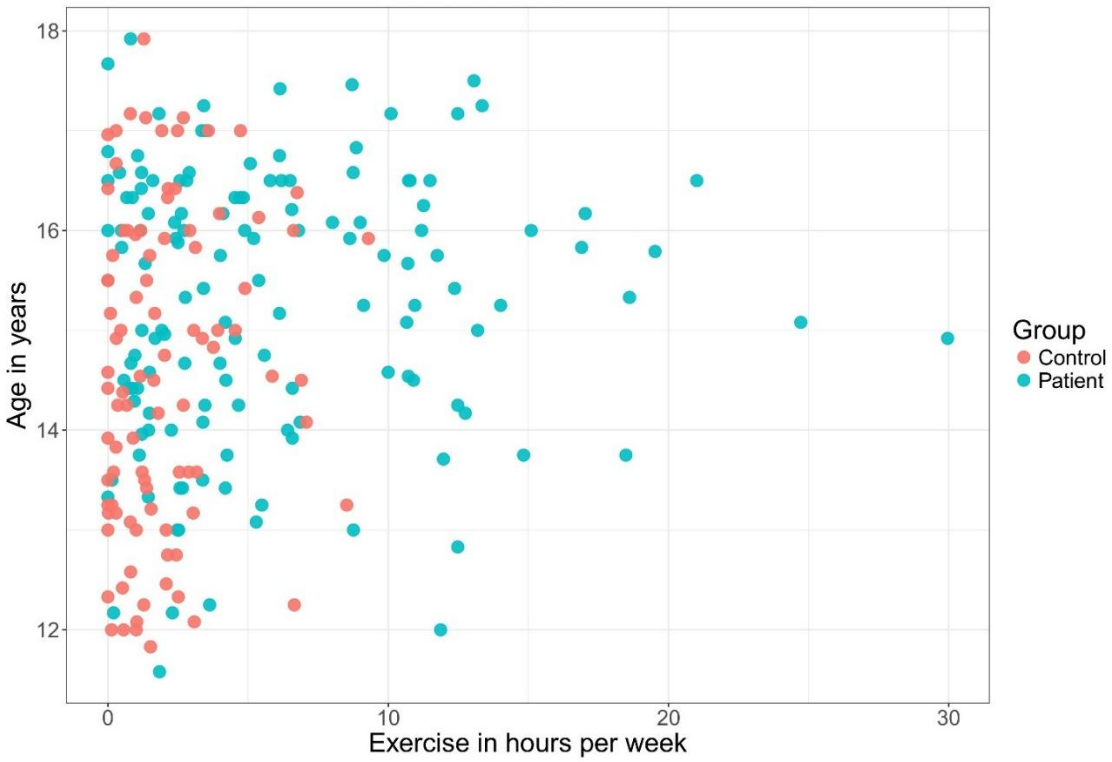


Figure 1: Observations before any splits.

The CART algorithm splits the data into boxes by using recursive binary splitting (Hastie et al., 2009). This approach is computationally greedy and top-down, meaning it starts with the entire data as well as all predictors and at each step the best split for that particular step is chosen, disregarding that this split might be suboptimal during later iterations (James, Witten, Hastie, & Tibshirani, 2013). This has downsides, but searching globally for a best tree that considers all splits at the same time is computationally infeasible (James et al., 2013). For classification, the split that best divides the observations is given by minimizing the Gini index:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Where $\hat{p}_{mk}$ is the proportion of observations in the *m*th box that would be formed by the split in the *k*th class (James et al., 2013). For the current example, this would mean minimizing the proportion of participants that are patients vs controls in a box given by a split. This index becomes smaller the "purer"

a box is, i.e. the more homogenous the participants inside are, and is thus referred to as a measure of node purity. For regression the analog to the Gini index is the residual sum of squares (RSS) given by:

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

Where $\hat{y}_{R_j}$ is the mean value of the dependent variable for participants within the *j*th box (James et al., 2013). This yet again is effectively a measure of variance inside the box produced by a given split and seeks to minimize that variance.

After a split has been made, the algorithm repeats the above process. For each split after the first however, only the regions yielded by prior splits are considered, because the alternative would be computationally infeasible. This means, that for the example given in figure 2, no split could ever cross the line of an earlier split. This procedure is repeated until a stopping criterion is met, such as that every new box needs to have more than five observations inside of it, in order to prevent overfitting. For the current example, the following first two splits are generated:

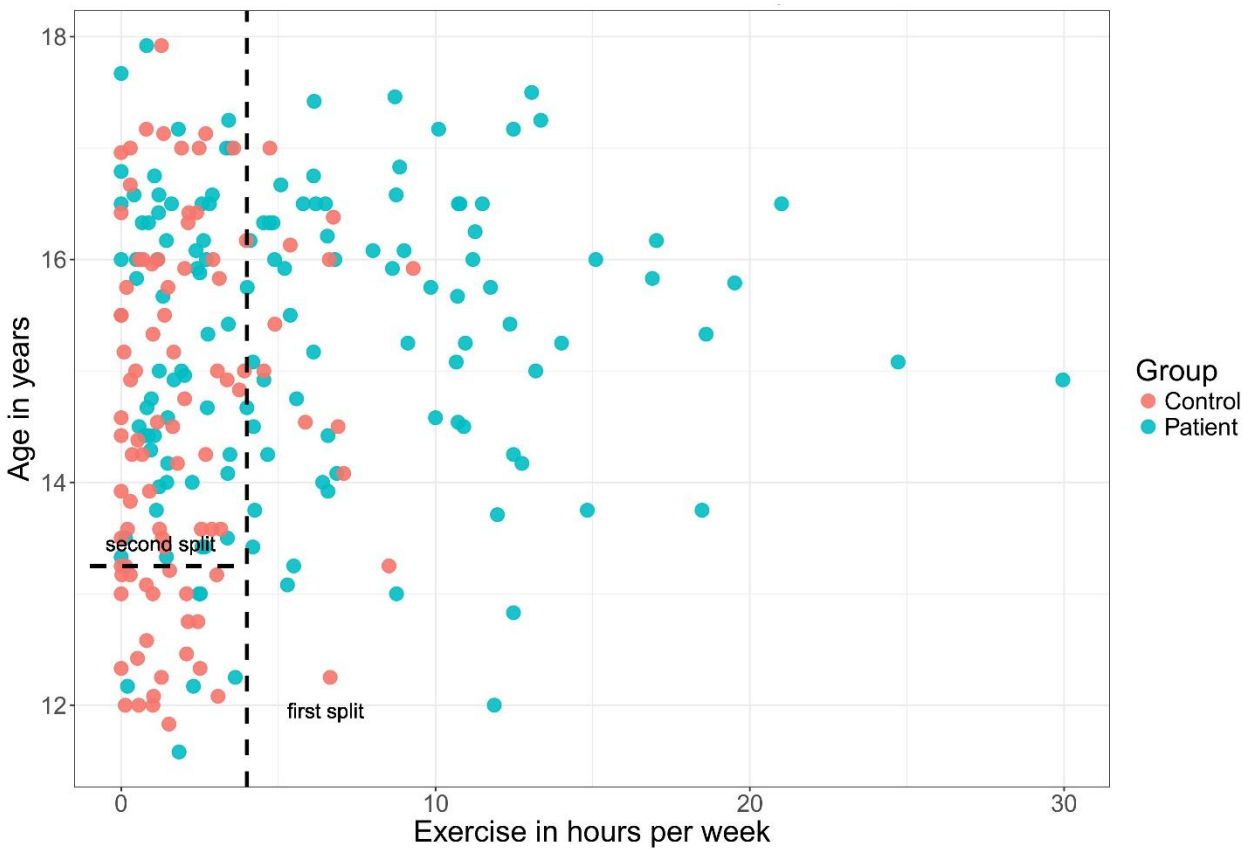


Figure 2: Observations after two splits.

In the current example with a two-dimensional variable space the splits can be visualized as simple lines dividing the variable space into two-dimensional boxes, but the concept remains the same in arbitrary many dimensions where the boxes become high-dimensional rectangles. The algorithm has found that a split on "Exercise in hours per week" at four hours best separates the groups, resulting in the observations to the right of the first split to be mostly patients. For the second split, the algorithm could choose "Exercise in hours per week" again, but has instead chosen the second independent variable "Age in years" at approximately 13.25 years. As stated above, the second split can only partition the space created by the first split, so the line for the second split stops at four on the "Exercise in hours per week" axis. The fact that the algorithm splits the variable space into boxes has several practical advantages for interpretation (James et al., 2013). Each tree can be visualized in the form of a decision tree that can be followed from top to bottom for each new observation that one wishes to classify.

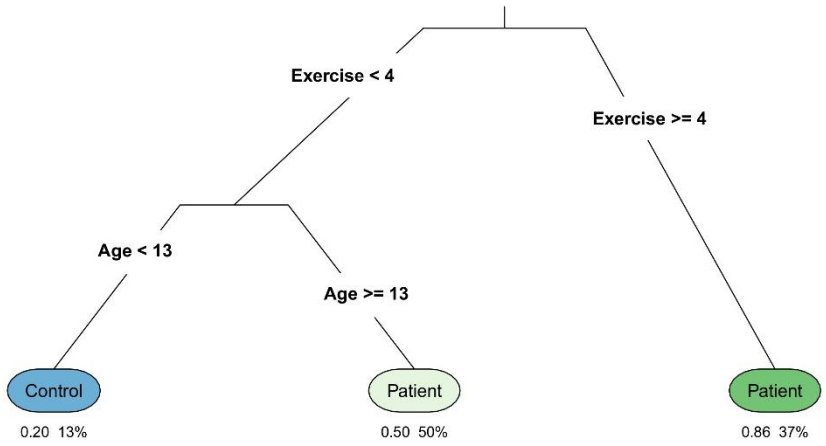


Figure 3: Decision tree stemming from the splits in figure 2.

The "leaves" at the bottom of the tree, which correspond to the boxes in figure 3 display the predicted class in the leaf, the predicted probability to be in that class, and the percentage of observations in that leaf. Every new observation will get dropped down the tree, and at each branch it is checked whether or not the observation satisfies one of the conditions (i.e. does the new participant exercise more or less than 4 hours a week?) corresponding to the splits in figure 2. The observation then ends up at one of the terminal leaves and is classified based on the majority class in the corresponding box. For example, if a new participant enters the study and exercises 2.5 hours a week and is 12 years old, following the decision tree above this would mean to go left on the first branch and left at the second branch, consequently predicting that the new participant belongs to the control group with a high certainty. For regression, the new observation would be predicted to have a value of the dependent variable equal to the mean value of all observation in the associated terminal leaf. This form of presentation of CART is easily interpretable as it provides clear cutoffs, but displays the associated uncertainty of these cutoffs in the form of error probabilities. Also, it grants an intuitive feeling for how important certain variables are by how far up in the tree they are or how prominently they show up in branches. At the same time, the decision tree diagram manages to visualize highly dimensional data and interactions clearly. The non-

parametric approach, which can often yield better prediction accuracy than classical regression models, along with the interpretability of the tree visualization are the prime features of CART. Some applications in psychotherapy research where these advantages have been applied to good use include studies about decision making in clinical practice (e.g. Hannöver, Richard, Hansen, Martinovich, & Kordy, 2002; Mann et al., 2008) or in search of treatment moderators for psychotherapy outcomes (see King & Resick, 2014 for an overview).

2.2. Random Forest

A disadvantage of CART is that it is not very robust, meaning that small changes in the data can cause both changes in variables on which the data is split, as well as where on those variables the data is split. In other words, CART suffers from high variance, which in turn leads to classification and mean squared error rates that are sub-par when compared with more state-of-the-art ML methods. One approach to improve the prediction accuracy of CART by reducing the variance is RF, invented by Breiman (2001a). A RF is an ensemble of classification or regression trees. The idea is to grow a forest out of many de-correlated trees, taking a vote of what the outcome should look like from every tree, and averaging the vote over all trees in the forest, thereby reducing the variance of the forest as a whole. The core change here is that the trees need to be de-correlated in order for the forest to work. In practice this is done by restricting the variables from which each tree is allowed to choose from at every split to a random subset m out of all available variables p. If a forest is grown without restricting the number of variables that can be chosen from at each split, the nature of the greedy algorithm used in CART would select the same strong predictor variables for all trees over and over again. Hence, there would be no advantage in averaging over many trees. This is can be suboptimal for prediction accuracy, because the greedy algorithm chooses the variable and cut-point to split only by taking in information from all previous splits without optimizing on splits yet to come. This means that, even though the variable chosen maximizes node purity at the current node, it might restrict the space in a way that is ultimately

not optimal for the entire tree (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008). In contrast, drawing a random set of predictors at each split means that in some splits the strongest predictors are not even considered and other predictors have more of a chance. This in turn makes the algorithm more robust to small changes in the data and as a result more reliable, while at the same time increasing prediction accuracy. Another advantage of the ensemble is that averaging over many trees smooths the hard decision boundaries that are created by splits in single trees (see also Biau, Devroye, & Lugosi, 2008).

2.3. Hyperparameters

RF has several so-called hyperparameters that can be adjusted to increase performance. One advantage of RF over other algorithms such as neural networks or GBM is that it has comparatively few hyperparameters and, for most problems, has quite good "out of the box" performance with all hyperparameters set at their default value (James et al., 2013; Strobl, Malley, & Tutz, 2009). Table 1 shows all hyperparameters for RF along with a brief description and typical default values for most software implementations of RF (Probst, Wright, & Boulesteix, 2019).

Table 1: Overview of the different hyperparameter of RF and typical default values

Hyperparameter	Description	Typical default values
mtry	Number of drawn candidate variables in each split	$\sqrt{p}$, p/3
Sample size	Number of observations that are drawn for each tree	n
Replacement	Draw observations with or without replacement	TRUE (with replacement)
Node size	Minimum number of observations in a terminal node	1 for classification, 5 for regression
Number of trees	Number of trees in the forest	500, 1000
Splitting rule	Splitting criteria to be optimized for homogenizing the nodes	Gini impurity, p value, random

Note. n = number of observations; p = number number of variables in the dataset. Table adapted with permission by authors (Probst, Wright, & Boulesteix, 2019)

A central hyperparameter of RF that can be highly relevant in problems typical for psychotherapy research, such as investigating the relevance of risk or resilience factors for a specific psychopathology, is

the number of variables drawn at each split, *mtry*. Typically, *mtry* is set to $mtry \approx \sqrt{p}$ and has shown to perform reasonably well with a variety of problems (Bernard, Heutte, & Adam, 2009). It can be tweaked by including more or less variables depending on the problem. In general, lower values of *mtry* result in less correlated trees than higher values of *mtry*. Higher values can in turn lead to higher stability when aggregating over the forest (Probst et al., 2019). However, lower values of *mtry* favor variables with moderate to low effects on the response that would otherwise be overshadowed by a few variables with a very strong effect on the response. These variables however, might be important to properly predict a subset of observations for which the stronger variables fail to predict correctly (Probst et al., 2019). Accordingly, lower values of *mtry* perform better in sets of variables with many variables of moderate effect on the predicted variable (Bernard et al., 2009). If, in contrast, the variable space is sparse, high values of *mtry* are preferable to ensure that one of the strong variables is found in each of the *mtry* sets for each split (Goldstein, Polley, & Briggs, 2011).

2.4. Variable importance

While CART are easily interpretable, RF is not. Pulling a single tree from a RF does not reveal everything about the RF as a whole. Due to the nature of the deliberate randomness introduced in the RF, some variables might not even show up. Hence, it is not feasible to simply draw an average tree from the RF to visualize for interpretation, as was done in the example above. Instead, because each variable gets a chance to be included in each split and therefore in different contexts, RF can provide much more nuanced measures of variable importance. These variable importance measures are especially important in cases where the aim is to explain a phenomenon rather than predicting it. In these scenarios they help to find the best predictors for a given response out of a broad set of candidate variables (Shmueli, 2010; Yarkoni & Westfall, 2017). A common variable importance measure for RF is the so called “permutation accuracy importance” (Strobl et al., 2008). In this, the values of a variable are randomly permutated in order to erase any systematic association of that variable with the response. The importance of that

variable is then measured as the difference in prediction accuracy before and after the permutation, averaged over all trees in the RF. Through this procedure, the variables with the strongest influence on the predictor in a variety of contexts and interactions, are assigned the highest variable importance. One of the strengths of this kind of variable importance measure is that it provides a combined measure of both individual as well as multivariate impact of each variable (Strobl et al., 2009) Also, it is found to be more efficient to detect predictors with interaction effects in highly correlated data than univariate screening (Verikas, Gelzinis, & Bacauskiene, 2011).

2.5. Gradient boosting machines

Like RF, GBM are also a form of ensemble learning, the difference being that the base learners in a GBM are trained sequentially. GBM are more complex than RF since they have more hyperparameters, which can be more powerful if calibrated well (e.g. Ogutu, Piepho, & Schulz-Streeck, 2011). Since GBM are much more complex, only a short conceptual idea is provided on the idea behind GBM since the algorithm was used alongside RF in the studies in this dissertation project. GBM, like RF is an approach that seeks to improve the prediction accuracy of models by reducing the variance. In contrast to RF, GBM can be applied to many statistical models. GBM starts with an initial model for the data, which in the case of the studies in this dissertation project, is a single tree with few splits. The algorithm then constructs a new model by successively fitting new base learners to the residuals of the current model rather than the outcome. The result of this new tree is then incorporated into the model to form a new iteration and the process is repeated for a set amount of iterations (James et al., 2013). The final GBM model is a linear combination of all trees that can be thought of a regression model with each tree representing one term in the regression equation (Elith, Leathwick, & Hastie, 2008). The GBM learns slowly by specifically targeting the areas of the data where the prior models do not do well, i.e. areas of the response variable for which the residuals are large. This process is further slowed down by a shrinkage parameter that shrinks the contribution of each individual tree to the final model. By fitting

small trees or other base learners to the residuals of the current model, the accuracy of the model is slowly improved for parts of the observations where it did not perform well prior. The most important hyperparameters for boosting models using trees are the shrinkage parameter *eta* and the interaction depth *max_depth*. The shrinkage parameter *eta* is a small positive number with typical values varying between 0.1 and 0.001. Smaller values of *eta* result in a slower learning rate and thus need more iterations to achieve good performance. The interaction depth *max_depth* controls the complexity of the boosted trees. An interaction depth of 1, corresponding to a decision tree with only two terminal nodes, fits an additive model while values of 2 and above fit models with two-way interactions and so on.

2.6. Applications

A recent review by Aafjes-Van Doorn and colleagues (2020) identified 51 psychotherapy research studies using ML. These 51 studies can be divided into 44 studies applying ML models primarily for data analysis and 7 studies reporting the feasibility of an ML assisted treatment tool. Of the 44 studies using ML models primarily for data analysis, 27 studies aimed at predicting the response of patients to an intervention either in terms of an outcome measure or drop-out. 12 Studies utilized ML in automated behavioral coding such as mimicking human raters in classifying linguistic categories. 7 Studies aimed at predicting process markers such as sudden gains from intake and outcome data. With regard to sample size, most study samples were of modest size with only 14 studies having samples of 200 or more participants. A wide variety of supervised and unsupervised ML algorithms were used. The most frequently used algorithms were artificial neural networks and support vector machines, being used in 11 studies each. 6 Studies used an RF based algorithm. In terms of validation procedures, 14 studies used leave-one-out CV, 13 studies used 10-fold CV, 7 studies used hold-out samples, 6 studies used other methods of validation and four did not describe any validation procedures.

An example of a study that used ML both for process and outcome prediction was conducted by Rubel and colleagues (2019). They used moderators of within person alliance-outcome associations of

already treated patients, as an example of a typical process-outcome association, in order to predict the process-outcome association for future patients. In the study, Rubel and colleagues used RF to select 11 variables out of a pool of 95, which best moderated the alliance-outcome association. Subsequently, they used a k-nearest neighbor algorithm (e.g. Hastie et al., 2009). The k-nearest neighbor algorithm selects patients from the training sample, which are similar in the previously identified process-moderator variables, and thus predicts the strength of the alliance-outcome association for new patients. These and similar approaches would enable clinicians to move from differential treatment selection to process or strategy selection tailored to the patient and informed by algorithms (e.g. Lutz, Zimmermann, Müller, Deisenhofer, & Rubel, 2017).

The study by Idalski Carcone and colleagues (2019) on the other hand, is an example of using ML, not to predict a clinical outcome, but instead training algorithms on transcripts classified by human experts to replicate their ratings. In this, they used transcripts from 37 sessions of motivational interviewing for weight loss with clinically obese patients, coded for 30 therapist and 16 patient behaviors, resulting in 11,353 coded utterances, and trained several different algorithms on the data. The best performing model in the training data, evaluated using 10-fold CV, was support vector machines (SVM; Cortes & Vapnik, 1995) scoring an average accuracy of 75.1%. The unmodified SVM model was then transferred to a HIV clinical care context and tested on 80 transcripts, achieving 72% accuracy. Automated classification using ML, as demonstrated in this study, have the potential to accelerate research using behavioral coding systems as they are much more efficient and faster compared to traditional human expert coding.

3. The present dissertation project

The aim of the present dissertation was to apply a variety of algorithms, both supervised and unsupervised, to a wide range of psychotherapy research problems, ranging from the search for risk and protective variables to predicting service utilization and aiding diagnostics, in order to best answer the underlying questions. By exploring the use of ML techniques such as RF and GBM as well as tests of generalization such as *k* fold CV, the author aims to contribute towards making these methods more understandable, familiar, and accessible to psychotherapy researchers. While advantages are highlighted, pitfalls and limitations of the method in general and the studies specifically are also outlined and discussed in detail.

4. Summary of the empirical studies

4.1. Study 1: One Size Fits All? Using Psychosocial Risk Assessments to Predict Service Use in Early Intervention and Prevention

Early intervention and prevention services in Germany offer a variety of different aides and interventions for families at risk. Interventions range from a parenting seminar as a form of universal prevention to home-visit programs as a form of selective prevention. However, there is often a mismatch between supply and demand with participants varying widely, and unexpectedly, in service utilization. Study one aimed at predicting participants' service use from routine screening data to better match participants to different interventions in a sample of N = 1,514 families at risk for negative childhood development. Routine screening data was utilized over additional questionnaires or interviews in order to help translating potential results into clinical practice.

The primary aim of the study was predicting the frequency of service utilization. Additionally the secondary aim of the study was to predict which kinds of intervention the participating families chose (out of medical services, social counseling, and counseling on the parent-child relationship) and how well the cooperation with the families was facilitated (dichotomized as either very well/well or rather

bad/bad). If the intervention was terminated on part of the participating family the reason for terminating the intervention was to be predicted (dichotomized as either regular end or other) as well as whether or not the participants were referred to other social services (dichotomized to additional referral or no referral). The outcome variables were based on the documentation done by the coordinators of the early prevention program while the 129 predictor variables were extracted from the risk screening tool (Heidelberger Belastungsskala, HBS, Sidor, Eickhorst, Stasch, & Cierpka, 2012) routinely employed by the home-visiting family nurses of the program. Several GBM classification (for all categorical outcomes) and regression (for frequency of service utilization) models were employed using CART as base learners. The hyperparameters were iteratively calibrated in training sets, utilizing 5-fold CV to avoid an overspill of information to the test set.

The final models applied to the test set were not successful in predicting the service utilization outcomes significantly above chance. Variable importance rankings identify socioeconomic risk variables as the only relevant predictors for all models.

While socioeconomic risk variables are important for the prediction of service utilization they do not appropriately reflect the spectrum of possible psychological reasons for service utilization. As a result, it is advised to extend the existing risk screening tool with variables accounting for constructs such as psychopathology, self-efficacy, attachment, and family functioning. While this study was sufficiently powered to support a GBM that was able to search for deep interactions in the variable space, no variables besides socioeconomic status emerged as significant predictors of service utilization. Future studies might profit from testing the performance of several different algorithms as alternatives to GBM as well as utilize screening tools with more psychologically minded items.

4.2. Study 2: Parenting stress in the face of early regulatory disorders in infancy – what matters most?

Early regulatory disorders (ERD) are the most prevalent diagnoses in under four year old children (Skovgaard et al., 2007) and are associated with high parental stress (Postert, Averbeck-Holocher, Achtergarde, Müller, & Furniss, 2012) . Risk and protective factors that are associated with ERD and might differentiate between families that cope well with the symptoms and those that experience high stress have yet to be explored. The aim of the second study was therefore to investigate the predictors of parenting stress in a sample of N = 135 mothers with infants diagnosed with early regulatory disorders (ERD) using a cross-sectional study design.

The sample was distinctively homogenous reflecting a sample with high socioeconomic status with 74.8% of mothers having achieved higher education and 79.3% of mothers being married, as well as 65.2% of the children being first born. The focus was on examining multiple factors from different realms. A multimethod approach including interviews and questionnaires was applied in order to account for different sources of information and to minimize the bias of self-reported data. Possible predictors of parenting stress (German Parenting Stress Index; Tröster, 2011) included psychological distress (Symptom-Checklist-90R-S; Franke, 2014), self-efficacy (Maternal Self-Efficacy Scale; Teti & Gelfand, 1991), parental reflective functioning (Parental Reflective Functioning Questionnaire; Ramsauer et al., 2014), infant development (Parent-Questionnaire, PQ; Cierpka, 2014), regulatory symptoms (Questionnaire for Crying, Feeding, and Sleeping; Gross, Reck, Thiel-Bonney, & Cierpka, 2013), socio-demographic data (PQ; Cierpka, 2014), and pre-, peri-, and postnatal risk factors for ERD (PQ; Cierpka, 2014), assessed using self-report questionnaires and behavioral diaries. Furthermore, a structured clinical interview was used to check whether or not the potential participants met the inclusion criteria for either persistent excessive crying, sensory processing, sleeping or feeding disorders. Additionally, those interviews were used to assess the parent-infant relationship, psychosocial risk, organic problems, and social-emotional functioning. From these measures a total of 464 variables were extracted and used

in the prediction of parenting stress by employing a GBM both with and without a feature selection algorithm. Recursive backwards selection of variables based on importance ranking out of the 464 variables was used to improve prediction performance. This resulted in 50 variables that were found to be the most informative ones in terms of outcome, and were included in the final model. The model was trained using 5-fold cross validation with 10 repeats in a test set of 70% of all observations and tested in a hold-out sample of the remaining 30% of observations.

The GBM algorithm with feature selection predicted parental stress with an RMSE of 21.72 and a mean absolute error of 17.04 which is within two thirds of the standard deviation (M = 131.5, SD = 31.60) of the outcome and performed significantly better than the model without feature selection ($t(15.3) = 3.4$; $p < .01$). The adjusted R^2 of the model was .58. The strongest predictors of parenting stress among the 464 variables were peripartum risk factors, and variables associated with the current problems in the mother-infant dyad (maternal symptoms of depression and irritability, maternal self-efficacy, and the regulatory symptoms of the child, especially fussing and crying.

The relatively small test set as well as the very specific nature of the sample characteristics prohibit a generalization of this study's findings to a broader, non-clinical context. Nevertheless, the accuracy achieved in this study is good both in the context of the outcome measures broad SD and range as well as the lack of knowledge on reliable predictors of parenting stress in this population. It stands to reason that one of the factors behind the success of the final model lies in the precursor variable selection algorithm. The RF based algorithm was able to reduce the variable set to 50 variables, thereby helping the subsequent GBM to focus on interactions and subgroups within the meaningful predictors, increasing the performance of all models with feature selection significantly. Going forward a replication of this study's findings is needed both because this is one of the first studies to explore factors related to parental stress in an ERD sample and to test, whether the factors found in this specific population can be generalized to other samples of families with ERD.

4.3. Study 3: Developing an assessment of epistemic trust: a research protocol / The Epistemic Trust Assessment (ETA) – An experimental measure of Epistemic Trust

The third study consists of two papers: The first paper titled “Developing an assessment of epistemic trust: a research protocol” provides a research protocol along with a theoretical framework for the conceptualization and development of an empirical assessment for ET, while the second paper “The Epistemic Trust Assessment (ETA) – An experimental measure of Epistemic Trust” describes the empirical study and pilot testing of the ETA.

ET can be described as the willingness to accept new interpersonally transmitted information as trustworthy, generalizable beyond the specific situation where it has been learned, and relevant to the addressee as an individual (Sperber et al., 2010). Contemporary psychodynamic theories have put forward the idea that a pervasive failure to establish ET might be the foundation of personality disorders (Fonagy, Luyten, & Allison, 2015). Furthermore, since the transmission of information is key to successful psychotherapies in general, it has been proposed that ET might be a working mechanism of all psychotherapies (Fonagy et al., 2015). Since there is currently no validated measure of ET available, this study set out to develop an experimental procedure with the aim of assessing ET in a non-clinical population.

The experiment was piloted in a sample of N = 61 university students. The Trier Social Stress Test for Groups (TSST-G, Dawans, Kirschbaum, & Heinrichs, 2011) was administered to the participants in an effort to induce stress and heighten relevance. The TSST-G involves a public speaking and mental arithmetic tasks in front of a committee of two evaluators and other experimental subjects. Heart rate monitors were set up so that the participants had reason to believe that the committee had assessed several physiological measures during the TSST-G. Afterwards, the computerized Epistemic Trust Questionnaire (ETQ) was administered in which the mock job interview portion of the TSST-G served as a standardized subject to give the participants feedback from the committee. The participants had to assess their performance during the TSST-G by both answering a yes or no question and rating their

certainty in statements in three different categories: Physiological (e.g. “Was your heartrate over or under 98?”), relational (e.g. “Do you think you came across as friendly?”), and mental-states (e.g. “Were you anxious?”). The participants then received feedback in which the evaluators were “trustworthy informants” (e.g., subjects’ objectively measured physiology: “Your heartrate during the interview was 120 bpm”), “untrustworthy informants” (e.g., subjects’ mental states: “The committee had the impression that you were anxious during the experiment”), or mix of both. Then, the participants had to re-rate the initial statements, having the opportunity to adjust their certainty ratings. The ET score was operationalized as the extent to which participants generalized the relevant feedback (e.g. physiological feedback, and relational feedback congruent with their own assessment), and rejected the irrelevant feedback (e.g. mental-states feedback and relational feedback incongruent with their own assessment). It was hypothesized that such a derived ET score would be approximately normally distributed in a healthy sample. Social desirability and PD traits were controlled for using the short scale for social desirability (KSE-G, Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt, 2012) and the Inventory of Personality Organization (IPO-16, Zimmermann et al., 2013). Additionally, an unsupervised agglomerative cluster analysis was employed to extract patterns of ET in the sample. Complete linkage was chosen as it avoids chaining problems encountered by single linkage approaches (Yim & Ramdeen, 2015) and Euclidean distance was used to compute the dissimilarity matrix.

The results of the study confirmed the hypothesis in that the participants, stemming from a non-clinical population, endorsed feedback relevant to them and rejected it otherwise, and that the resulting ET score had an approximate normal distribution. With regards to the cluster analysis, three clusters of participants were found in the sample termed “overly vigilant”, “naïve/-uncertain”, and “adaptive”. The overly vigilant subgroup was characterized by relatively high initial certainties as well as little change post feedback. The naïve/-uncertain group was characterized by low initial certainty in self-states, as well as high change in certainty in the self-states category. The adaptive group had low initial certainties in

physiological as well as relational states, as well as high change in the physiological and relational categories.

These findings closely resemble the hypothetical clusters described in the literature (e.g. Fonagy & Allison, 2014; Sperber et al., 2010). From this pilot study the conclusion could be drawn that the ETA can be used as the first internally validated measure of ET. However, the ETA still has to be externally validated in a clinical population.

4.4. Study 4: Assessing Personality Functioning and Maladaptive Traits in Young Adults: A Machine Learning Approach

There are several current theories of PDs, that individually all lack scope, comprehensiveness and most importantly, empirical support (e.g.; Karterud & Kongerslev, 2019). While the rationale for new, emerging models such as the alternative model of the DSM-5 (AMPD; American Psychiatric Association, 2013) and the ICD-11 (Tyrer, Reed, & Crawford, 2015) have gathered empirical support (Zimmermann, Kerber, Rek, Hopwood, & Krueger, 2019), several issues remain. Validated measures are extensive and show little agreement between different methods and data sources (Oltmanns & Oltmanns, 2019). In study four, the use of ML for the prediction of categorical and dimensional PD as well as maladaptive personality traits was evaluated, with the aim of achieving valid and reliable diagnostics as a byproduct of routine outcome measurement.

Categorical and dimensional diagnosis of PD were derived from the Levels of Personality Functioning Scale – Self Report (Morey, 2017) representing criterion A, while maladaptive traits, representing criterion B, were assessed using the Personality Inventory for DSM-5 Brief Form (American Psychiatric Association, 2013). Additionally, the study aimed at identifying patterns in variables commonly associated with PD such as attachment, mentalizing, childhood trauma, interparental conflict and parental rejection, with the goal of deriving data driven predictors of PDs to compare with theoretically derived predictors listed in the literature. To this end a GBM with CART as base learners

was trained in a non-clinical sample of 410 young adults. The data was split 70% into a training set and 30% into a test set. Hyperparameters for the GBM were calibrated by iterating over possible combinations of shrinkage between 0.01 and 0.2, the interaction depth of each tree between 1 and 6, the number of boosting iterations between 1 and 1000. These values represent a cautious approach with little risk of overfitting and a focus on sensitivity over specificity as it was deemed more important to reliably identify all individuals with PD than to correctly classify all individuals without PD. All other hyperparameters were fixed at their respective default values.

For the prediction of the primary outcome, categorical PD, a prediction accuracy of 91.06% was achieved at a NIR of 85.37% (*p* accuracy > NIR = 0.042) with a sensitivity of 95.24% and a specificity of 66.67%. For dimensional personality functioning, an accuracy of RMSE = 46.10 was reached, which corresponds to 67% of the standard deviation for personality functioning in the sample or 13.9% of the observed range. The R^2 for the model was 0.57. In terms of predicting maladaptive personality traits, sufficient accuracy for detachment, psychoticism, and negative affect were achieved but not for antagonism and disinhibition. The most important variables for the prediction of a present diagnosis of PD, as well as maladaptive traits, was both attachment avoidance and attachment anxiety, both measured with the Experiences in Close Relationships – Revised (ECR-RD; Ehrenthal, Dinger, Lamla, Funken, & Schauenburg, 2009). Values of avoidance above 60 and anxiety above 70 seemed to indicate a much higher chance of being diagnosed with a PD.

The results of the study are promising since individuals with a PD were identified based on peripheral variables only. Additionally, the study provides empirical support for the novel Temperament-Attachment-Mentalization Theory (TAM; Karterud & Kongerslev, 2019), as well as mentalization based theories of PD (Fonagy et al., 2015). This study showcases the potential for ML to be utilized in diagnostics. This is especially relevant in diagnostics of phenomena like PD where, while there are several competing theories on PD, no established consensus exists on the factors which underpin PD. With instruments that assess PD being complex and lengthy, ML in this study was able to identify a

pattern of peripheral variables that yet yields very high sensitivity towards detecting PD at 95.24%. Nevertheless, since the sample of this study did not consist of inpatients, the study would have to be replicated in a clinical sample.

4.5. Study 5: The Impact of Trainee Attributes and Training Variables on Competence Deterioration: Results from a Longitudinal Study in Naturalistic German Psychotherapy Training

While emerging research on the effects of psychotherapy training show positive effects of training on some trainee competences (Willutzki, Fydrich, & Strauß, 2015), few studies have investigated whether or not subgroups of trainees deteriorate during training and what risk- or protective-factors might be associated with these subgroups. The fifth study was set out to both quantify and predict deterioration of personal and professional variables of trainees undergoing the German psychotherapy training.

The study used data from a German study on trainee development (Evers & Taubner, 2019) that followed N = 184 trainees over a timespan of three years. Systematic deterioration was operationalized as reliable deterioration from the pre-assessment to the post-assessment on at least one of the following scales: Healing Involvement, Stressful Involvement and Basic Relational Skills (Work Involvement Scales), Attributional Complexity (Attributional Complexity Scale), and Introject Affiliation (Intrex Questionnaire). The scales were chosen to cover a variety of core professional and personal competencies that could be described as the “outcome” of psychotherapy training. 52.31% of trainees fulfilled the criteria for systematic deterioration. Following, a conditional inference RF algorithm was utilized using a variety of variables covering 5 domains: childhood trauma, attachment strategies, professional background, personality traits, life satisfaction, therapeutic attitude, training aspects, training context, and sociodemographic variables. 5-fold CV with 10 repeats was used to assess the generalizability of the results. As a consequence of not having a separate test-set of observations to independently assess

generalizability, the algorithm used was not calibrated in its hyperparameters but instead used default values and one model only to avoid overfitting.

The RF algorithm achieved an average overall accuracy of 67.54%, a specificity of 66.62%, and a sensitivity of 69% averaged over 10 repeats of 5-fold CV. The most important domain was life satisfaction, which when combined with attachment strategies, was highly indicative of deterioration.

These findings from a naturalistic setting highlight the need to put more emphasis on routinely monitoring negative outcomes in psychotherapy training, and to have measures in place in case vulnerable trainees deteriorate. While the ML algorithm applied in this study was successful in achieving a significantly higher accuracy than the NIR, and provides insight into the variables predicting deterioration, it is nevertheless questionable whether or not the achieved accuracy, specifically the sensitivity, is high enough for any real-world applications. In order to further raise prediction accuracy more variables that might be linked to deterioration, or more cases to be analyzed are needed.

5. Discussion

Overall, all five empirical research studies utilized ML algorithms in an effort to help guide and inform theories and practices.

In the first study, we aimed at predicting participant behavior in service utilization from routine screening data. Service utilization behavior was measured in terms of several categorical, as well as continuous outcomes. For the categorical outcomes, it was not possible to train an algorithm that successfully predicted any of the service utilization behaviors significantly better than by classifying every family according to the most common class. For example, while a classification accuracy of 84.04% was achieved for predicting cooperation behavior, an accuracy of 86.17% was reached by classifying every family as having “good” cooperation. Likewise, for the regression models, good accuracy in terms of small RMSE was not attained. While the results of the prediction approach were comparable (Brand & Jungmann, 2014), or better (Daro, McCurdy, Falconnier, & Stojanovic, 2003; Goyal et al., 2016) than

other studies in the field, they are nevertheless lackluster for practical purposes. A possible reason for this could be that the GBM was not able to find the combination of items that would have successfully predicted the outcomes and other algorithms might have done better. However, GBM is regarded as one of the best ML algorithms for data with predefined items (e.g. Fan et al., 2018; Ogutu et al., 2011). Also, to the best of the authors knowledge there is no comparable study using ML to predict service utilization in families at risk for negative childhood development. While Brand and Jungmann (2014) achieved similar results using logistic regression in a sample of N = 434 socially disadvantaged mothers, who enrolled in an early intervention and prevention program, there are several core differences in study methodology. While study one focused on data from a routine risk-screening measure with no additional assessments, Brand and Jungmann (2014) had data available both from additional face-to-face interviews with the participants as well as additional questionnaires covering psychosocial constructs in far more depth. Also, the logistic regression models utilized by Brand and Jungmann (2014) were trained on the entire sample with no estimate of sample generalizability, and thus overfitting, supplied. In sum, it is difficult to judge whether the ML approach has merit for this research question because it is unclear if the result can be attributed to the specific algorithm used, shortcomings of the screening measure, or a combination of both. Going forward more studies are needed and studies should utilize a variety of algorithms and regression approaches along with a modified risk screening measure.

The second study aimed to gain insight into the complex interactions of variables that dictate whether or not parents experience heightened parental stress in a sample of families with early regulatory disorders. To this end, parenting stress was predicted from a set of 464 variables using GBM. The final algorithm attained a RMSE of 21.72, when applied to the test set. To the best of the author's knowledge, this study is the first to have explored factors related to parenting stress in ERD by including multiple measures and searching for interactions. The study's findings suggest that maternal self-efficacy in combination with exhaustion and the duration of infant fussing and crying are the most important out of the 464 variables with regard to predicting parenting stress. Additionally, the ability of the ML

algorithm to model non-linear relationships indicates that all three of the aforementioned variables have an incremental “threshold” before they have an effect on parenting stress. This indicates that, for example, maternal self-efficacy values in the upper middle range do not have a negative effect on parenting stress compared to values in the higher range, but self-efficacy values below the middle range have an abrupt and pronounced effect on parenting stress. A similar effect emerges for exhaustion and infant fussing and crying. Summarizing, this study highlights the uses of ML in being able to model non-linear relationships and interactions with several different predictors in the same model. Because of the homogeneous and specific sample of highly educated first mothers, the results of this study have to be replicated in a separate sample. However, the results of this study, similar to other ML applications in the study of risk and protective variables (e.g. Ribeiro et al., 2019), might move the field forward and inform new treatment approaches.

The third study aimed at piloting an experimental paradigm to measure ET in a sample of university students. Agglomerative hierarchical clustering was utilized, resulting in three mostly distinct clusters. Interestingly, the attained clusters revealed subtle patterns close to what is hypothesized in the literature for PD samples, albeit less marked. If replicated in a larger sample, the results of the cluster analysis might be very important for research on ET, as almost all of the theoretical literature describing the working mechanisms and patterns of ET are centered on individuals with PDs and little is known about healthy samples.

In the fourth study, a GBM was trained to predict both categorical and continuous measures of personality functioning, as well as maladaptive traits, in a sample of 410 young adults based on a number of variables regarded as risk factors for the development of PD. Both continuous as well as categorical PD were predicted well. Additionally, the prediction of negative affect, detachment, and psychoticism was successful. However, good prediction accuracy was not achieved for antagonism and disinhibition. While it is difficult to properly assess the accuracy of the algorithms concerning the continuous measures because of a lack of similar studies, the results are nevertheless promising. With regard to categorical

measures, the achieved sensitivity of 95.24% lies markedly above the average sensitivity of 80% achieved by a variety of PD measures in a meta-analytic review (Gárriz & Gutiérrez, 2009), while the specificity of 66.67% achieved by the GBM is lower than the average specificity of 73%. While, to the best knowledge of the authors, this is the first study using ML to help PD diagnostics the study corroborates results from prior studies in psychotherapy research using ML to replicate human rating or aid with diagnostics. For example Hatton and colleagues (2019) were able to use a GBM to aid with the diagnosis of persistent depressive symptoms from auxiliary variables, achieving an accuracy of 89%. The results of this study lend strong support to the notion that ML might be used as a tool in diagnostics for PD. Going forward, if the results of this study can be replicated in clinical populations using state of the art PD interviews as learning data, ML algorithms might be able to assist costly, long diagnostic interviews for the diagnosis of PD. In a psychiatric context, this could be used to first administer a short battery of questionnaire items based on the results of this study and only assess potential PD candidates using a time-consuming interview measure if the algorithm finds evidence for a possible PD. This would save significant resources on the side of the clinic, as PD diagnostic interviews can take up to two hours, while also being more patient friendly than conventional diagnostics. This approach, however, is only viable because of the high sensitivity yielded by the algorithm, and the validation procedure which indicates that the found pattern is not specific to the training sample but can be generalized to a wider population.

In the fifth study, an RF algorithm was employed in order to predict who would and would not deteriorate of the 184 psychotherapy trainees with the aim of identifying prognostic variables. The results showed that the algorithm successfully predicted 67.54% of the participants correctly over 10 repeats of 5-fold CV. In the future, calibrating the algorithm for a high sensitivity is more important than a high specificity since the goal is to identify potential deteriorating candidates as early as possible, with the aim of helping them more adaptively cope with the stress of psychotherapy training.

In terms of sample sizes, the dissertation showed that ML can be used in a range of differently sized samples. However, as shown in the first study, larger N do not automatically translate into higher

accuracies. What does however benefit greatly from larger sample sizes are the validation procedures. Larger sample sizes both enable dedicated hold-out sets for validation in combination with *k*-fold CV validated training within the training set, as well as more complex algorithms such as neural nets or GBM.

Of the algorithms used in this dissertation, the RF algorithm used in the fifth study stood out as being easy to use because of it's relatively low number of hyperparameters. This favors RF in small sample studies over algorithms with more calibration potential such as neural nets, as these typically perform worse without extensive iterative calibration, which in turn can easily lead to overfitting in small samples (Strobl et al., 2009). However, since this also transfers, to a degree, into being less flexible when adapting to different problems RF might not achieve the highest accuracies in studies with larger sample sizes (Strobl et al., 2009). Additionally, conditional variable importance measures are available for RF which provide variable importance without being biased towards correlated variables, which is a common problem for other algorithms (Strobl et al., 2008).

5.1. Limitations

While some of the results of ML applications highlighted in this dissertation are promising, some fall short. There are several limitations of this project that make it difficult to answer the question of what the causes for the poor performance displayed by some models in this project might have been. Most of the studies utilized small samples for ML techniques. This in turn restricted the amount of model calibration that could be done without running the risk of overfitting. Without being able to run several different algorithms on the same data, the question which algorithm is best fit for which research design cannot be answered. The sample sizes also affect the ability to translate the results of the project into clinical practice. The results of study two, for example, could be utilized in the treatment of ERD by focusing on restoring a medium level of maternal self-efficacy and prevent high levels of exhaustion as the association pattern found suggests that this is more effective in reducing parenting stress than

focusing on one resilience factor alone. However, as the non-linear association pattern might be an artifact of overfitting to the specific sample, bigger data sets along with the stronger validation procedures are needed to enable ML to more easily transfer research results into clinical practice. Future studies should strive to pool samples from similar populations, for example from different outpatient departments that treat ERD, in order to take full advantage of ML algorithms. Finally, another limitation of this project that goes along with the novelty of ML applications in psychotherapy research is the lack of similar studies to accurately put the results in context. This is of course not only a limitation for the present study but for most ML applications in psychotherapy research that cannot define a clear target accuracy from theoretical assumptions. For example, in the fifth study, if an intervention program was available to address trainees at risk for deterioration, a screening algorithm's desired minimal accuracy might have been informed by the cost of the intervention. Also, results from studies utilizing ML cannot be easily compared with prior studies, since studies using linear regression variants rarely report estimates of generalizability. In study 4, for example, we achieved an RMSE of 46.10 with an adjusted R^2 of 0.57 when predicting dimensional personality functioning. While the R^2 value indicates that a large proportion of the variance in dimensional personality functioning has been explained by the model, without further studies to compare these results to, it remains a judgement call whether or not this prediction accuracy could be helpful in any clinical application.

5.2. Future Directions

This dissertation has shown that ML algorithms using pre-defined variables, such as questionnaire items, or scores from observational measures, can answer psychotherapy research questions such as identifying risk- and protective variables. Additionally, there is vast potential for psychotherapy research in algorithms that do not use pre-defined input variables, but instead find predictive patterns on their own. Future studies might explore the application of algorithms such as convolutional neural nets that are already used in diagnostic applications such as computer vision

(Bernal et al., 2019) or voice recognition (Nassif, Shahin, Attili, Azzeh, & Shaalan, 2019) outside of psychotherapy research. These algorithms could be utilized as tools, enabling better workflow, or even enable research designs that were either previously not possible or not feasible. For example, these algorithms might be used in the future to enable quicker workflow by automatically transcribing audio or video files. This would enable large scale studies using complex observational measures that were not previously feasible. Expanding this further, neural nets could be trained on transcripts, based on the premise that the researcher is able to provide ratings of those transcripts for the training set, to also automate the rating process for future transcripts of the same measure. Researchers would only have to code a minor portion of the transcripts themselves to prove the reliability of the algorithm. Psychotherapy process research is another area that would benefit immensely from the use of such algorithms, as it often involves complex behavioral coding instruments that are traditionally associated with high workloads for transcription and coding. Additionally, areas such as synchrony research (Delaherche et al., 2012) and emotion recognition and regulation (Healy, Donovan, Walsh, & Zheng, 2018) are on the forefront of utilizing ML techniques to advance our understanding of how psychotherapy works and for whom.

5.3. Conclusion

In this dissertation, ML algorithms were described in terms of their basic underlying principles and employed to answer several different research questions. These research questions could not have been answered by classical statistical models such as logistic regression or multiple regression without either violating their statistical assumptions (e.g. multicollinearity) or risking non-convergence. However, it is important to state that the research question should always dictate the method used and not the other way around. Consequently, ML algorithms should be understood as an extension to the statistical toolkit that can be employed to answer questions of psychotherapy research. By providing supplemental data and analysis code where possible (e.g. study 3 and the examples in this work), the

authors aim to contribute towards ML algorithms being more accessible. Overall, this dissertation provides an overview and some examples on how ML can be considered a valuable tool for exploring complex multivariate data. These techniques being popularized both in psychology in general and psychotherapy research specifically, offer a promising future direction.

6. References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2020). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, 1–25. https://doi.org/10.1080/10503307.2020.1808729

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (5th ed.).*

Bernal, J., Kushibar, K., Asfaw, D. S., Valverde, S., Oliver, A., Martí, R., & Lladó, X. (2019). Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: A review. *Artificial Intelligence in Medicine*, *95*, 64–81. https://doi.org/10.1016/j.artmed.2018.08.008

Bernard, S., Heutte, L., & Adam, S. (2009). Influence of Hyperparameters on Random Forest Accuracy. In J. A. Benediktsson, J. Kittler, & F. Roli (Eds.), *Lecture Notes in Computer Science. Multiple Classifier Systems* (Vol. 5519, pp. 171–180). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02326-2_18

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of Random Forests and Other Averaging Classifier. *Journal of Machine Learning Research*, *9*, 2015–2033.

Brand, T., & Jungmann, T. (2014). Participant characteristics and process variables predict attrition from a home-based early intervention program. *Early Childhood Research Quarterly*, *29*(2), 155–167. https://doi.org/10.1016/j.ecresq.2013.12.001

Breiman, L. (1998). *Classification and regression trees* (Repr). Boca Raton: Chapman & Hall.

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical Modeling:: The Two Cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cierpka, M. (2014). *Frühe Kindheit 0-3 Jahre: Beratung und Psychotherapie für Eltern mit Säuglingen und Kleinkindern* (2., korr. Aufl.). Berlin: Springer. Retrieved from http://dx.doi.org/10.1007/978-3-642-39602-1 https://doi.org/10.1007/978-3-642-39602-1

Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Daro, D., McCurdy, K., Falconnier, L., & Stojanovic, D. (2003). Sustaining new parents in home visitation services: Key participant and program factors. *Child Abuse & Neglect*, *27*(10), 1101–1125. https://doi.org/10.1016/j.chiabu.2003.09.007

Dawans, B. von, Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology*, *36*(4), 514–522. https://doi.org/10.1016/j.psyneuen.2010.08.004

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines. *IEEE Transactions on Affective Computing*, *3*(3), 349–365. https://doi.org/10.1109/T-AFFC.2012.12

Ehrenthal, J. C., Dinger, U., Lamla, A., Funken, B., & Schauenburg, H. (2009). Evaluation der deutschsprachigen Version des Bindungsfragebogens "Experiences in Close Relationships--Revised" (ECR-RD) [Evaluation of the German version of the attachment questionnaire "Experiences in Close Relationships--Revised" (ECR-RD)]. *Psychotherapie, Psychosomatik, medizinische Psychologie*, *59*(6), 215–223. https://doi.org/10.1055/s-2008-1067425

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/ j.1365-2656.2008.01390.x

Evers, O., & Taubner, S. (2019). Kompetenzentwicklung in der Psychotherapieausbildung. *PiD - Psychotherapie Im Dialog*, *20*(04), 58–63. https://doi.org/10.1055/a-0771-7912

Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., . . . Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management*, *164*, 102–111. https://doi.org/10.1016/j.enconman.2018.02.087

Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic Petrification and the Restoration of Epistemic Trust: A New Conceptualization of Borderline Personality Disorder and Its Psychosocial Treatment. *Journal of Personality Disorders*, *29*(5), 575–609. https://doi.org/10.1521/pedi.2015.29.5.575

Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology*, *87*(8), 684–692. https://doi.org/10.1037/ccp0000421

Franke, G. L. (2014). *SCL-90®-S. Symptom-Checklist-90®-Standard–Manual*: Hogrefe.

Franklin, J. C. (2019). Psychological primitives can make sense of biopsychosocial factor complexity in psychopathology. *BMC Medicine*, *17*(1), 187. https://doi.org/10.1186/s12916-019-1435-1

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gárriz, M., & Gutiérrez, F. (2009). Cribado de trastornos de la personalidad: un metaanálisis [Personality disorder screening: a meta-analysis]. *Actas espanolas de psiquiatria*, *37*(3), 148–152.

Goldstein, B. A., Polley, E. C., & Briggs, F. B. S. (2011). Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, *10*(1), 32. https://doi.org/10.2202/1544-6115.1691

Goyal, N. K., Folger, A. T., Hall, E. S., Teeters, A., van Ginkel, J. B., & Ammerman, R. T. (2016). Multilevel assessment of prenatal engagement in home visiting. *Journal of Epidemiology and Community Health*, *70*(9), 888–894. https://doi.org/10.1136/jech-2014-205196

Gross, S., Reck, C., Thiel-Bonney, C., & Cierpka, M. (2013). Empirische Grundlagen des Fragebogens zum Schreien, Füttern und Schlafen (SFS) [Empirical basis of the Questionnaire for Crying, Feeding and Sleeping]. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *62*(5), 327–347. https://doi.org/10.13109/prkk.2013.62.5.327

Hannöver, W., Richard, M., Hansen, N. B., Martinovich, Z., & Kordy, H. (2002). A Classification Tree Model for Decision-Making in Clinical Practice: An Application Based on the Data of the German Multicenter Study on Eating Disorders, Project TR-EAT. *Psychotherapy Research*, *12*(4), 445–461. https://doi.org/10.1080/713664470

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hatton, C. M., Paton, L. W., McMillan, D., Cussens, J., Gilbody, S., & Tiffin, P. A. (2019). Predicting persistent depressive symptoms in older adults: A machine learning approach to personalised mental healthcare. *Journal of Affective Disorders*, *246*, 857–860. https://doi.org/10.1016/j.jad.2018.12.095

Healy, M., Donovan, R., Walsh, P., & Zheng, H. (2018). A Machine Learning Emotion Detection Platform to Support Affective Well Being. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* Madrid, Spain.

Huang, X., Ribeiro, J. D., & Franklin, J. C. (2019). *The Differences between Suicide Ideators and Suicide Attempters: Simple, Complicated, or Complex?* https://doi.org/10.31234/osf.io/8tuqg

Idalski Carcone, A., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., . . . Kotov, A. (2019). Developing Machine Learning Models for Behavioral Coding. *Journal of Pediatric Psychology*, *44*(3), 289–299. https://doi.org/10.1093/jpepsy/jsy113

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E., & Steinley, D. (2020). *Pairing Machine Learning and Clinical Psychology: How You Evaluate Predictive Performance Matters.* https://doi.org/10.31234/osf.io/2yber

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Karterud, S. W., & Kongerslev, M. T. (2019). A Temperament-Attachment-Mentalization-Based (TAM) Theory of Personality and Its Disorders. *Frontiers in Psychology*, *10*, 518. https://doi.org/10.3389/fpsyg.2019.00518

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G). Retrieved from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-339589

Kendler, K. S. (2019). From Many to One to Many-the Search for Causes of Psychiatric Illness. *JAMA Psychiatry.* Advance online publication. https://doi.org/10.1001/jamapsychiatry.2019.1200

King, M. W., & Resick, P. A. (2014). Data mining in psychological treatment research: A primer on classification and regression trees. *Journal of Consulting and Clinical Psychology*, *82*(5), 895–905. https://doi.org/10.1037/a0035886

Lee, S. H., Maenner, M. J., & Heilig, C. M. (2019). A comparison of machine learning algorithms for the surveillance of autism spectrum disorder. *PloS One*, *14*(9), e0222907. https://doi.org/10.1371/journal.pone.0222907

Lutz, W., Zimmermann, D., Müller, V. N. L. S., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: Study protocol. *BMC Psychiatry*, *17*(1), 306. https://doi.org/10.1186/s12888-017-1464-2

Mann, J. J., Ellis, S. P., Waternaux, C. M., Liu, X., Oquendo, M. A., Malone, K. M., . . . Currier, D. (2008). Classification trees distinguish suicide attempters in major psychiatric disorders: A model of clinical decision making. *The Journal of Clinical Psychiatry*, *69*(1), 23–31. https://doi.org/10.4088/jcp.v69n0104

Martinez, J., Black, M. J., & Romero, J. (2017). On Human Motion Prediction Using Recurrent Neural Networks. *IEEE*, 4674–4683. https://doi.org/10.1109/CVPR.2017.497

Morey, L. C. (2017). Development and initial evaluation of a self-report form of the DSM-5 Level of Personality Functioning Scale. *Psychological Assessment*, *29*(10), 1302–1308. https://doi.org/10.1037/pas0000450

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access*, *7*, 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, *5 Suppl 3*, S11. https://doi.org/10.1186/1753-6561-5-S3-S11

Oltmanns, J. R., & Oltmanns, T. (2019). *Self–Other Agreement on Ratings of Personality Disorder Symptoms and Traits: Three Meta-Analyses.* https://doi.org/10.31234/osf.io/mka3j

Postert, C., Averbeck-Holocher, M., Achtergarde, S., Müller, J. M., & Furniss, T. (2012). Regulatory disorders in early childhood: Correlates in child behavior, parent-child relationship, and parental mental health. *Infant Mental Health Journal*, *33*(2), 173–186. https://doi.org/10.1002/imhj.20338

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), 281. https://doi.org/10.1002/widm.1301

R Development Core Team (2017). *R: A language and environment for statistical computing*: Vienna. Retrieved from http://www.R-project.org

Ramsauer, B., Lotzin, A., Mühlhan, C., Romer, G., Nolte, T., Fonagy, P., & Powell, B. (2014). A randomized controlled trial comparing Circle of Security Intervention and treatment as usual as interventions to increase attachment security in infants of mentally ill mothers: Study Protocol. *BMC Psychiatry*, *14*, 24. https://doi.org/10.1186/1471-244X-14-24

Ribeiro, J. D., Huang, X., Fox, K. R., Walsh, C. G., & Linthicum, K. P. (2019). Predicting Imminent Suicidal Thoughts and Nonfatal Attempts: The Role of Complexity. *Clinical Psychological Science : A Journal of the Association for Psychological Science*, *7*(5), 941–957. https://doi.org/10.1177/2167702619838464

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., & Lutz, W. (2019). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches-A demonstration. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, 1–10. https://doi.org/10.1080/10503307.2019.1597994

Schmitgen, M. M., Niedtfeld, I., Schmitt, R., Mancke, F., Winter, D., Schmahl, C., & Herpertz, S. C. (2019). Individualized treatment response prediction of dialectical behavior therapy for borderline personality disorder using multimodal magnetic resonance imaging. *Brain and Behavior*, *9*(9), e01384. https://doi.org/10.1002/brb3.1384

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sidor, A., Eickhorst, A., Stasch, M., & Cierpka, M. (2012). Einschätzung der Risikobelastung in Familien im Rahmen von Frühen Hilfen: Die Heidelberger Belastungsskala (HBS) und ihre Gütekriterien [Assessing risk exposure in families within the scope of early intervention: the Heidelberg Stress scale (HBS) and its quality criteria]. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *61*(10), 766–780. https://doi.org/10.13109/prkk.2012.61.10.766

Skovgaard, A. M., Houmann, T., Christiansen, E., Landorph, S., Jørgensen, T., Olsen, E. M., . . . Lichtenberg, A. (2007). The prevalence of mental health problems in children 1(1/2) years of age - the Copenhagen Child Cohort 2000. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *48*(1), 62–70. https://doi.org/10.1111/j.1469-7610.2006.01659.x

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25*(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Teti, D. M., & Gelfand, D. M. (1991). Behavioral competence among mothers of infants in the first year: The mediational role of maternal self-efficacy. *Child Development*, *62*(5), 918–929. https://doi.org/10.1111/j.1467-8624.1991.tb01580.x

Tröster, H. (2011). *Eltern-Belastungs-Inventar: EBI; German version of the Parenting Stress Index (PSI) of R. R. Abidin*: Hogrefe.

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet*, *385*(9969), 717–726. https://doi.org/10.1016/S0140-6736(14)61995-4

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, *44*(2), 330–349. https://doi.org/10.1016/j.patcog.2010.08.011

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science : A Journal of the Association for Psychological Science*, *5*(3), 457–469. https://doi.org/10.1177/2167702617691560

Willutzki, U., Fydrich, T., & Strauß, B. (2015). Aktuelle Entwicklungen in der Psychotherapieausbildung und der Ausbildungsforschung. *Psychotherapeut*, *60*(5), 353–364. https://doi.org/10.1007/s00278-015-0048-1

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yim, O., & Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology*, *11*(1), 8–21. https://doi.org/10.20982/tqmp.11.1.p008

Zimmermann, J., Benecke, C., Hörz, S., Rentrop, M., Peham, D., Bock, A., . . . Dammann, G. (2013). Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). *Diagnostica*, *59*(1), 3–16. https://doi.org/10.1026/0012-1924/a000076

Zimmermann, J., Kerber, A., Rek, K., Hopwood, C. J., & Krueger, R. F. (2019). A Brief but Comprehensive Review of Research on the Alternative DSM-5 Model for Personality Disorders. *Current Psychiatry Reports*, *21*(9), 92. https://doi.org/10.1007/s11920-019-1079-z

7. Acknowledgements

I want to thank all of you who have supported me throughout this process and who stood beside me during this journey in this time of my life.

First, I want to thank my mentor Svenja Taubner who has supported me ever since my very first, slightly out of the ordinary, job application by showing me the scientific world and letting me take part in such events as an exclusive tour through old Jerusalem or Athens.

Second, I owe gratitude to my wonderful colleagues, Anna, Oli and Max, who have been my main partners in crime and who have supported me not only professionally but who have also been there for me personally, every step of the way.

Third, I feel grateful for all the wise advisers in the form of Jana, Alessandro, Markus, and Steffi who I was able to ask at any time and that would always, despite their extremely busy schedules find time for my inquiries.

Last but not least, I feel honored to call such excellent human beings such as Regina and Horst my in-laws as well as my parents, Diane and Klaus, my wonderful wife Ann-Christin and my son Alexander. I cannot count the hours in which you had my back because I had to work yet another weekend on my dissertation and do not know what I would have done without your seemingly unending well of patience and compassion. You are both my source of inspiration and strength.

8. Declaration in accordance to § 8 (1) c) and (d) of the doctoral degree regulation of the Faculty

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

FAKULTÄT FÜR VERHALTENS- UND EMPIRISCHE KULTURWISSENSCHAFTEN

Promotionsausschuss der Fakultät für Verhaltens- und Empirische Kulturwissenschaften der Ruprecht-Karls-Universität Heidelberg
Doctoral Committee of the Faculty of Behavioural and Cultural Studies of Heidelberg University

Erklärung gemäß § 8 (1) c) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften
Declaration in accordance to § 8 (1) c) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation selbstständig angefertigt, nur die angegebenen Hilfsmittel benutzt und die Zitate gekennzeichnet habe.
I declare that I have made the submitted dissertation independently, using only the specified tools and have correctly marked all quotations.

Erklärung gemäß § 8 (1) d) der Promotionsordnung der Universität Heidelberg für die Fakultät für Verhaltens- und Empirische Kulturwissenschaften
Declaration in accordance to § 8 (1) d) of the doctoral degree regulation of Heidelberg University, Faculty of Behavioural and Cultural Studies

Ich erkläre, dass ich die vorgelegte Dissertation in dieser oder einer anderen Form nicht anderweitig als Prüfungsarbeit verwendet oder einer anderen Fakultät als Dissertation vorgelegt habe.
I declare that I did not use the submitted dissertation in this or any other form as an examination paper until now and that I did not submit it in another faculty.

Vorname Nachname
First name Family name
Paul Schröder-Pfeifer

Datum, Unterschrift
Date, Signature
12.11.2020,

9. Appendix

Study 1

I. Evers, O., & **Schröder, P.** (2018). One Size Fits All? Using Psychosocial Risk Assessments to Predict Service Use in Early Intervention and Prevention/One size fits all? Die Eignung von Risikoscreenings zur Prognose der Inanspruchnahme von Angeboten der Frühen Hilfen. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *67*(5), 462-481. IF: 0.513

Declaration of author contributions: Evers, O.: Conceptualization, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Schröder, P.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Methodology, Software, Formal analysis

One size fits all? Die Eignung von Risikoscreenings zur Prognose der Inanspruchnahme von Angeboten der Frühen Hilfen

Oliver Evers und Paul Schröder

Summary

One Size Fits All? Using Psychosocial Risk Assessments to Predict Service Use in Early Intervention and Prevention

Early intervention and prevention services offer a variety of programs. At the same time, program participants differ widely in their service use. This study aims at investigating the prognostic validity of psychosocial risk assessments in predicting the participants' service use. The psychosocial risk assessment "Heidelberg Stress Scale" is used to predict aspects of service use (dosage, attrition, intervention content, working relationship). Service use data of N = 1.514 participants of a home-visiting program will be analyzed via Machine-Learning-Algorithms. Dosage and intervention content can be predicted with psychosocial risk assessments. The classification strength is small. Global and continuous risk scales have a prognostic advantage over single categorical risk items. Financial burden has a significant influence on every aspect of service use. Psychosocial risk assessments provide additional information that can support intervention planning. Yet, these instruments should be supplemented by additional diagnostic information.

Prax. Kinderpsychol. Kinderpsychiat. 67/2018, 462-480

Keywords

risk assessment – home visiting – service use – predictive validity – psychosocial risk

Zusammenfassung

Frühe Hilfen haben eine heterogene Angebotsstruktur. Familien unterscheiden sich zudem in der Inanspruchnahme der Angebote. Es wird untersucht, inwiefern psychosoziale Risikoscreenings verschiedene Aspekte der Inanspruchnahme vorhersagen. Inanspruchnahmevariablen (Nutzungsintensität, Beendigung, Interventionsinhalte, Weitervermittlung, Arbeitsbündnis) sollen durch das Risikoscreening „Heidelberger Belastungsskala" vorhergesagt werden. Dazu werden Inanspruchnahmedaten von N = 1.514 Teilnehmern eines Hausbesuchsprogramms mithilfe von Machine-Learning-Algorithmen untersucht. Nutzungsintensität und Interventionsinhalte lassen sich durch psychosoziale Risikoscreenings vorhersagen. Die Klassifikationsgüte ist jedoch eingeschränkt. Numerische Gesamtrisiko-Einschätzungen sind dabei prognostisch wichtiger als kategoriale Einzelfaktoren. Finanzielle Belastungen

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

https://doi.org10.13109/prkk.2018.67.5.462

haben einen bedeutsamen Einfluss auf alle Aspekte der Inanspruchnahme. Psychosoziale Risikoscreenings können im Sinne einer differenziellen Indikationsstellung zusätzliche Informationen zur Interventions- und Fallzahlplanung liefern. Sie sollten jedoch nicht als einziges diagnostisches Instrument eingesetzt werden.

Schlagwörter

Frühe Hilfen – psychosoziales Risiko – Inanspruchnahme – Risikoscreening – prognostische Validität

1 Hintergrund

Die Frühen Hilfen zeichnen sich durch eine große Vielfalt an Angeboten und Zielgruppen aus. Gleichzeitig weisen erste Erfahrungen auf eine heterogene Inanspruchnahme durch verschiedene Klientengruppen hin. In der vorliegenden Arbeit soll eine Möglichkeit untersucht werden, das Inanspruchnahmeverhalten von Klienten mithilfe von psychosozialen Risikoscreenings zu prognostizieren und deren Eignung für eine differenzielle Indikationsstellung auszuloten.

1.1 Interventionsformen und Inanspruchnahme der Angebote in den Frühen Hilfen

In der Modell- und Implementierungsphase der Frühen Hilfen wurden verschiedene Interventionsformen entwickelt oder an den deutschen Sprachraum adaptiert (Cierpka u. Evers, 2015). Die Interventionen beinhalten Angebote im Bereich der universellen Prävention (Cierpka, Gregor, Frey, 2004), der selektiven Prävention für Familien mit multiplen psychosozialen Risikofaktoren (Brand u. Jungmann, 2010; Cierpka, 2009; Suess, Bohlen, Mali, Frumentia Maier, 2010; Ziegenhain, 2007) und der indizierten Prävention (Wiegand-Grefe, Halverscheid, Plass, 2011). In der Implementierungsphase der Frühen Hilfen etablierten die meisten Kommunen nicht nur eine spezifische Interventionsform, sondern eine breite Palette an parallelen Hilfsangeboten (Nationales Zentrum Frühe Hilfen [NZFH], 2014).

Zusätzlich fällt eine hohe Heterogenität in der Nutzung einzelner Interventionen auf. So setzen die meisten selektiven Präventionsangebote für psychosozial belastete Familien in ihrem Konzept thematische Schwerpunkte. In der Praxis wird jedoch berichtet, dass Familien die Intervention für eine Vielzahl von Anliegen in Anspruch nehmen. In qualitativen Interviews untersuchten Kitzman, Cole, Yoos und Olds (1997) die Implementierung eines amerikanischen Hausbesuchsprogramms. Die 17 befragten Fachkräfte berichteten vor allem von Hindernissen, die Programmziele der elterlichen Gesundheit und der kindlichen Entwicklung zu thematisieren. Die Fachkräfte brachten dies mit akutem Bedarf für Sozialberatung, der Koordination mit ver-

schiedenen Helfern und der Prävalenz von Familienkonflikten in Verbindung. Eine ähnliche Tendenz zeigt sich in der Prozessevaluation des deutschen Programms „Pro Kind". Hier wurde deutlich, dass die Beschäftigung mit Aspekten der Elternrolle leicht unterrepräsentiert war, während multiple gesundheitliche, soziale und kindliche Themen in den Hausbesuchen in den Vordergrund traten (Brand u. Jungmann, 2010).

Die Aufgabe der Zuweisung innerhalb der heterogenen Angebotsstrukturen übernehmen häufig Koordinationsstellen in den Netzwerken der Frühen Hilfen (NZFH, 2014) oder einzelne Netzwerkakteure (vgl. Künster, Knorr, Fegert, Ziegenhain, 2010). Die Anpassung der Interventionsinhalte obliegt häufig den durchführenden Fachkräften. Eine Hilfestellung zur differenziellen Zuweisung und Anpassung der Programminhalte könnte eine systematische Einschätzung der Familien anhand von objektivierbaren Merkmalen bieten. Einige Merkmale werden in den Frühen Hilfen bereits großflächig mit Instrumenten erfasst, die im nächsten Abschnitt vorgestellt werden.

1.2 Systematische Erfassung des psychosozialen Risikos

Psychosoziale Risikoscreenings dienen der systematischen Erfassung von Merkmalen, die in empirischen Studien mit einer erhöhten Wahrscheinlichkeit für eine negative kindliche Entwicklung oder mit einem Misshandlungsrisiko in Verbindung gebracht wurden. Die Angemessenheit von Risikoscreenings wurde innerhalb der Frühen Hilfen ausgiebig diskutiert (Kindler, 2010). Kenntnis und Anwendung dieser Screeninginstrumente hat mittlerweile Eingang in die Kompetenzprofile von aufsuchenden Fachkräften gefunden (Hahn u. Sandner, 2013). Zudem berichtete im Jahr 2014 die Hälfte der Kommunen, standardisierte Instrumente zur Beurteilung von psychosozialen Risikovariablen zu nutzen (NZFH, 2014).

In systematischen Übersichten werden bis zu 27 verschiedene Instrumente zur Erfassung von psychosozialem Risiko genannt (van der Put, Assink, Boekhout van Solinge, 2017). Zu empirisch entwickelten Instrumenten in Deutschland zählen das Kurzscreening „Anhaltsbogen für ein vertiefendes Gespräch" (Kindler, 2009) und die ausführlichere „Heidelberger Belastungsskala" (HBS; Sidor, Eickhorst, Stasch, Cierpka, 2012). Aktuelle psychosoziale Risikoscreenings basieren auf der Einschätzung von Misshandlungsrisiko und kindlichen Entwicklungsrisiken (Übersicht in Bender u. Lösel, 2014). Damit zielen sie auf eine allgemeine Indikationsstellung für den Interventionsbedarf (vgl. Kindler, 2010), nicht jedoch auf eine differenzielle Indikation, welche Angebote angezeigt sind. Für eine Einschätzung der differenziellen Indikation sind Erkenntnisse nötig, inwiefern psychosoziale Risiken mit Aspekten des Bedarfs und der Angebotsnutzung in Verbindung gebracht werden können.

1.3 Prädiktoren für eine differenzielle Inanspruchnahme der Frühen Hilfen

Beim ersten Schritt der Inanspruchnahme, dem *Zugang* zu Interventionsangeboten, wurden Unterschiede bezüglich psychosozialer und medizinischer Belastung darge-

stellt. Psychosoziale Risikofaktoren wie familiärer Stress (Duggan et al., 2000) und Anzahl der Belastungsfaktoren (McCurdy et al., 2006) haben in bisherigen Studien zu einer häufigeren Teilnahme an selektiven Präventionsprogrammen geführt. Migrationserfahrungen der Eltern führten zu einer geringeren Teilnahme (Moore et al., 2005), Ergebnisse zu Alter und Bildungsgrad der Mütter unterscheiden sich nach Programmschwerpunkt (Duggan et al., 2000; Goyal et al., 2016). Medizinische Risikofaktoren der Kinder, wie ein niedriges Geburtsgewicht oder Frühgeburtlichkeit, waren in der Regel mit einer erhöhten Zugangsrate zu selektiven Präventionsprogrammen assoziiert (Duggan et al., 2000; McCurdy et al., 2006).

Unterschiede in der *Nutzung* der Intervention beziehen sich vor allem auf die Frequenz der Hausbesuche. Familien mit jüngeren oder arbeitslosen Müttern nahmen weniger Hausbesuche während der Programmlaufzeit in Anspruch (Daro, McCurdy, Falconnier, Stojanovic, 2003; Goyal et al., 2016). In einer Untersuchung psychologischer Eigenschaften der Mütter konnten Olds und Korfmacher (1998) zudem eine höhere Zahl an Hausbesuchen für Teilnehmerinnen mit einer geringen internalen Kontrollüberzeugung zeigen.

Beim *Interventionsabbruch* lassen sich soziodemografische und psychologische Risikovariablen unterscheiden. Soziodemografische Risikovariablen wie Minderjährigkeit der Mutter, häufige Umzüge und eine hohe Gewaltrate am Wohnort, werden mit einem Abbruch der Intervention in Verbindung gebracht (Brand u. Jungmann, 2014; Fraser, Armstrong, Morris, Dadds, 2000; McGuigan, Katzev, Pratt, 2003). Derweil sind psychologische und familiäre Risikovariablen wie Schwierigkeiten in der Eltern-Kind-Beziehung, geringe elterliche Selbstwirksamkeitserwartung, Depressivität, väterliche Stressbelastung sowie gegen die Mütter gerichtete Partnerschaftsgewalt prädiktiv für einen längeren Verbleib in der Intervention (Duggan et al., 2000; Fraser et al., 2000; Girvin, DePanfilis, Daining, 2007).

In den dargestellten Untersuchungen zeigt sich ein Trend zur Berücksichtigung von sozialen und medizinischen Risikofaktoren. Psychische Risikovariablen waren in bisherigen Inanspruchnahmestudien unterrepräsentiert, obwohl es Hinweise für eine zentrale Bedeutung in der Vorhersage von Interventionsabbrüchen gibt (Duggan et al., 2000; Fraser et al., 2000; Girvin et al., 2007). Weitere Limitation der bisherigen Studien sind die Berücksichtigung von einzelnen ausgewählten Prädiktoren und limitierten Inanspruchnahmevariablen, insbesondere dem Programmzugang und dem Interventionsabbruch. Bisher existiert jedoch keine Studie, die alle in der Praxis erfassten Risikobereiche in der Vorhersage von mehreren Aspekten der Inanspruchnahme berücksichtigt hat.

1.4 Aktuelle Studie und Fragestellung

Die dargestellten Studien unterstreichen die breite Angebotsstruktur und heterogene Inanspruchnahme von Programmen der Frühen Hilfen. In dieser Angebotslandschaft könnten objektivierbare Familienmerkmale eine Hilfestellung für die sy-

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

stematische Zuweisung zu passenden Angeboten bieten. Eine sparsame Möglichkeit zu deren Erhebung wäre die Nutzung von psychosozialen Risikovariablen, die in der Praxis bereits erfasst werden.

In der vorliegenden Studie sollen daher gesundheitliche, interaktionelle, soziale und familiäre Risikofaktoren betrachtet werden. Die Variablen werden in einer explorativen Fragestellung dahingehend untersucht, ob sich psychosoziale Risikoscreenings zur Vorhersage einer differenziellen Inanspruchnahme eignen, aus denen eine differenzielle Indikation abgeleitet werden kann. Zusätzlich sollen diejenigen Risikofaktoren identifiziert werden, die für die Prognose der Inanspruchnahme bedeutsam sind. Daraus ergeben sich folgende explorative Fragen:

1. Sind psychosoziale Risikoscreenings geeignete Instrumente zur Prognose einer differenziellen Inanspruchnahme?
2. Welche Aspekte der Inanspruchnahme können gut vorhergesagt werden?
3. Welche Risikovariablen dienen als gute Indikatoren?

In dieser Studie wird dafür auf Inanspruchnahmedaten des Hausbesuchsprogramms „Keiner fällt durchs Netz“ (KfdN) zurückgegriffen. Unter den Teilnehmern von KfdN werden Ergebnisse eines Risikoscreenings mit Angaben zur Nutzungsintensität (Anzahl der Hausbesuche, Anzahl der Netzwerkkontakte zu anderen Institutionen), Inhalte (medizinisch, Sozialberatung, Eltern-Kind-Interaktion), Beendigung (regulär, Abbruch), weiterem Bedarf (Weitervermittlung) und Arbeitsbündnis (Zusammenarbeit) in Zusammenhang gebracht werden.

2 Methode

Bei dieser Studie handelt es sich um eine explorative Analyse von Inanspruchnahmedaten der Hausbesuchsintervention des Programms „Keiner fällt durchs Netz“ (KfdN). Die Studie wurde in Übereinstimmung mit der Deklaration von Helsinki durchgeführt und durch die Ethikkommission der medizinischen Fakultät der Universität Heidelberg genehmigt. Die Teilnehmer des Programms haben der pseudonymisierten Verwendung der während der Intervention erhobenen Daten zugestimmt.

2.1 Das Programm „Keiner fällt durchs Netz“

KfdN wurde als Modellprojekt der Frühen Hilfen in den Jahren 2007 bis 2013 in neun deutschen Landkreisen eingeführt. Das Programm beinhaltet eine Netzwerk- und eine Interventionskomponente (Cierpka, 2009). Mit der Netzwerkkomponente wurden Koordinationsstellen und interdisziplinäre Netzwerke der Frühen Hilfen etabliert. Die Intervention besteht aus einem Elternkurs als universelle Präventionsmaßnahme (Cierpka et al., 2004) und einem Hausbesuchsprogramm durch aufsuchende Helferinnen als selektive präventive Intervention.

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

Die aufsuchenden Helferinnen im Hausbesuchsprogramm waren Familienhebammen, Sozialmedizinische Assistentinnen und Sozialpädiatrische Familienbegleiterinnen, die in einem 160-stündigen Curriculum geschult wurden. Der vorgesehene inhaltliche Schwerpunkt der Hausbesuchsintervention lag in der Förderung von elterlichen Kompetenzen und einer feinfühligen Eltern-Kind-Interaktion.

2.2 Rekrutierung und Stichprobe

Der Zugang der Familien zum Hausbesuchsprogramm von KfdN erfolgte freiwillig über die Empfehlung von Institutionen im Netzwerk Frühe Hilfen. Das Screening auf Einschlusskriterien und die Zuweisung der Familien erfolgte über die kommunalen Koordinationsstellen der Frühen Hilfen. Einschlusskriterien waren das Kindesalter (< 12 Monate), Wohnort in der Projektregion und das Vorhandensein von mindestens einem psychosozialen Risikofaktor. Bei erfolgreicher Programmvermittlung wurden zu Beginn der Hausbesuche ausführliche Risikoscreenings durch die aufsuchenden Helferinnen durchgeführt.

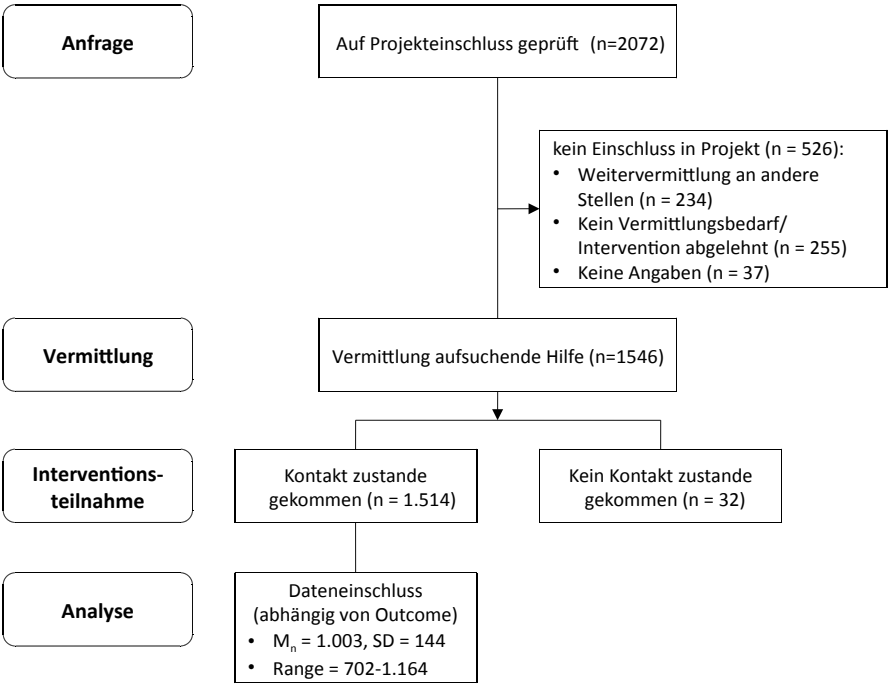


Abbildung 1: Probandenfluss

Der Probandenfluss ist in Abbildung 1 dargestellt. Für den Projektzeitraum von 2010 bis 2013 liegen Daten zu n = 2.072 Anfragen bei den Koordinierungsstellen der Frühen Hilfen vor. Davon wurden n = 1.546 (74,6 %) an aufsuchende Helferinnen vermittelt, wobei in n = 32 Fällen (2,1 %) kein Kontakt mit der Helferin zustande kam. Aus der endgültigen Stichprobe von n = 1.514 Interventionsfamilien wurden jeweils nur die Fälle in die Auswertung aufgenommen, bei denen die untersuchte Inanspruchnahmevariable vorlag. Daher variieren die Unterstichproben für die Analyse zwischen n = 702 und n = 1.164 (M_n = 1.003, SD = 144).

Für die Interventionsfamilien wurden 31,4 % der Anfragen während der Schwangerschaft (Schwangerschaftswoche: M = 29,44; SD = 7,96) und 68,6 % der Anfragen nach der Geburt des Kindes gestellt (Alter in Wochen: M = 6,70; SD = 9,03). Zuweisende Institutionen waren Nachsorgehebammen (31,9 %), Geburts- und Kinderkliniken (20,9 %), das Jugendamt (17,5 %), andere Einrichtungen aus dem sozialen Bereich (11,2 %) oder aus dem Gesundheitsbereich (9,9 %) sowie Selbstmelder (8,6 %).

2.3 Instrumente

2.3.1 Risikovariablen

Heidelberger Belastungsskala (HBS; Sidor et al., 2012): Die Langversion der HBS wurde zur Erfassung einzelner Risikofaktoren sowie zur summativen Einschätzung des Gesamtrisikos durch die aufsuchenden Helferinnen verwendet. In der HBS werden 124 einzelne Risikovariablen auf einer dreistufigen Skala eingeschätzt (trifft zu, Anzeichen, trifft nicht zu). Die Risikovariablen sind in die Bereiche „Belastung des Kindes", „Elterliche/Familiäre Belastung", „Soziale Belastung" und „Materielle Belastung" unterteilt. Für diese Kategorien sowie für die „Gesamtbelastung" wird zudem eine Gesamteinschätzung auf der Skala von 0 bis 100 vorgenommen. Nach Sidor et al. (2012) weist die HBS eine hohe Interraterreliabilität unter Ratern der gleichen Profession auf (ICC = 0,81-0,90), während die Interraterreliabilität zwischen Familienhebammen und geschulten Psychologiestudierenden niedrig ist (ICC = 0,21). Die prädiktive Validität für Fälle von Kindeswohlgefährdung liegt im mittleren Bereich (Sensitivität = 0,78; Spezifität = 0,74) und ist vergleichbar mit anderen Risikoscreenings (van der Put et al., 2017).

2.3.2 Inanspruchnahmevariablen

Falldokumentation: Zu jeder Anfrage bei den Koordinierungsstellen der Frühen Hilfen erfassten die Koordinatorinnen kurze Angaben zu Anliegen, Zuweisern und vorhandener Unterstützung. Bei erfolgreicher Vermittlung wurden im Fallverlauf Inanspruchnahmedaten wie die Zahl der Hausbesuche, Kontakte innerhalb des Frühe Hilfen Netzwerks, Weitervermittlungen und Beendigungsgründe dokumentiert.

Hausbesuchsdokumentation: Jeder Hausbesuch wurde durch die aufsuchenden Helferinnen in einem strukturierten Dokumentationsbogen zusammengefasst und Informationen zu Dauer, Anwesenden sowie besprochenen Inhaltsbereichen gegeben.

Abschlussdokumentation: Zum Abschluss der Intervention machten die aufsuchenden Helferinnen in einem kurzen Fragebogen Angaben zum Interventionsverlauf. Dabei schätzen sie in Bezug auf das Arbeitsbündnis die Zusammenarbeit mit der Familie, Zuverlässigkeit, Engagement und Umsetzbarkeit der Interventionsinhalte ein. Ebenso wurde die Kooperation mit anderen Einrichtungen beurteilt.

2.4 Statistische Analyse

2.4.1 Untersuchungsvariablen

Als *Prädiktoren* wurden alle 124 Einzelrisikofaktoren sowie die fünf Skalen zur Gesamtrisikoeinschätzung der HBS herangezogen. Für die Auswahl des *Kriteriums* wurde auf Variablen aus der bisherigen Literatur zurückgegriffen sowie auf weitere Variablen, die eine Aussage über eine differenzielle Inanspruchnahme erlauben. Untersucht wurden Angaben zur Nutzungsintensität (Anzahl der Hausbesuche, Anzahl der Netzwerkkontakte zu anderen Institutionen), zu Inhalten (medizinische Versorgung, Sozialberatung, Eltern-Kind-Interaktion), zur Beendigung (Regulär, Abbruch), zum weiteren Bedarf (Weitervermittlung) und Arbeitsbündnis (Zusammenarbeit).

2.4.2 Fehlende Werte

Die Daten enthielten zwischen 10,1 und 16,7 % fehlende Werte (M = 14.7 %; SD = 1.9 %). Die Imputation wurde unter der Missing-at-Random-Annahme (MAR) vorgenommen. Zur Imputation wurde Multiple Imputation by Chained Equations unter vollständig konditionaler Spezifikation mit 40 Imputations-Iterationen verwendet.

2.4.3 Analyseverfahren

Die prädiktiven Eigenschaften der erhobenen Risikovariablen wurden mittels Machine Learning (ML) Verfahren unter der Verwendung von Gradient Boosting Machines (GBM) mit Regressionsbäumen beurteilt. Für dichotome Outcomes wurde das finale Modell nach der höchsten Klassifizierungsgenauigkeit ausgesucht, für numerische Outcomes nach dem niedrigsten Root Mean Squared Error (RMSE). Vorgehen und Analyseparameter werden ausführlich in den Online-Materialien beschrieben.

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

3 Ergebnisse

3.1 Deskriptive Ergebnisse

Tabelle 1 gibt einen Überblick der deskriptive Daten zu den Inanspruchnahmevariablen. *Nutzungsintensität:* Die *Anzahl der projektfinanzierten Hausbesuche* lag im Mittel bei 18,81 Hausbesuchen (SD = 14,78). Die minimale Hausbesuchszahl von 0 ergibt sich für Familien, die bereits in der Regelversorgung von Familienhebammen betreut wurden und darüber hinaus keine Besuche wahrnahmen. Im Durchschnitt nahmen die Koordinatorinnen oder aufsuchenden Helferinnen im Interventionsverlauf mit 1,40 anderen Stellen (SD = 1,56) Kontakt auf.

Inhalt der Hausbesuche: Für Angaben zu Inhalten von Hausbesuchen wurden alle dokumentierten Hausbesuche der teilnehmenden Familien zusammengefasst. Dazu wurde jeweils ein Index gebildet, der den Anteil der Hausbesuche anzeigt, in dem ein bestimmter Themenbereich behandelt wurde. Am häufigsten wurden Themen der *medizinischen Versorgung* (z. B. Wochenbettbetreuung, medizinische Komplikationen, Versorgung des Kindes) behandelt (76,7 %). Themen der *Sozialberatung* wurden in 61,9 % der Hausbesuche besprochen. Der Hauptfokus des Programms, die *Eltern-Kind-Interaktion* (z. B. Wahrnehmung und Interpretation von kindlichen Signalen, Kommunikation, Entwicklungsberatung) wurde in 61,1 % der Hausbesuche thematisiert.

Tabelle 1: Deskriptive Statistiken der Inanspruchnahmevariablen

Inanspruchnahme		M	SD	Min	Max	n
Inhalt[1]	Eltern-Kind-Interaktion	0,611	0,323	0,00	1,00	981
	Sozialberatung	0,619	0,333	0,00	1,00	961
	Medizinische Versorgung	0,767	0,259	0,00	1,00	986
Anzahl	Netzwerkkontakte	1,40	1,56	0,85	6,00	1.102
	Hausbesuche	18,81	14,78	0,00	133,00	997
		Anteil %				**n**
Zusammenarbeit	Sehr gut / gut	83,5 %	-	-	-	702
	Eher schlecht / schlecht	16,5 %	-	-	-	
Weitervermittlung	Erfolgt	58,9 %	-	-	-	1.164
	Nicht erfolgt	41,1 %	-	-	-	
Beendigung	Regulär	78,1 %	-	-	-	1.132
	Nicht regulär	21,9 %	-	-	-	

Anmerkungen. [1]Anteil der Hausbesuche, in denen der jeweilige Inhalt behandelt wurde auf einer Skala von 0 bis 1

Beendigungsgründe wurden dichotomisiert zu regulären (z. B. Kindesalter von 12 Monaten, Erreichung von Interventionszielen) und nicht regulären (z. B. mangelnde Zusammenarbeit, Inobhutnahme, fehlende Passung der Intervention) Beendigungen. Der Großteil der Betreuungen (78,1 %) wurde regulär beendet.

Weitervermittlungen: Über die Hälfte der Familien (58,9 %) wurden im Laufe der Intervention an eine weitere Stelle vermittelt.

Arbeitsbündnis: In einer dichotomisierten Einschätzung des Arbeitsbündnisses schätzen die aufsuchenden Helferinnen die *Zusammenarbeit mit der Familie* überwiegend als sehr gut oder gut ein (83,5 %).

3.2 Vorhersagegüte der psychosozialen Risikoscreenings für Inanspruchnahmevariablen

Zur Beurteilung der Vorhersagegüte der Risikoscreenings wird die Genauigkeit der statistischen Modelle beurteilt, die auf den 124 Risikovariablen und 5 Risikoskalen der HBS beruhen. Die Vorhersagegenauigkeit dichotomer Inanspruchnahmevariablen wird dabei mit der No-Information-Rate (NIR) verglichen. Die NIR gibt an, wie genau das Vorhersagemodell wäre, wenn alle Familien der Kategorie mit der höchsten Häufigkeit zugeordnet würden. Wie in Tabelle 2 dargestellt, liegt die Vorhersagegenauigkeit für *Weiterverweisungen* bei 63,32 %, für die reguläre *Beendigung* der Intervention bei 83,89 % und für die *Güte der Zusammenarbeit* bei 84,04 %. Damit ist die Vorhersage jeweils besser als das Zufallsniveau. Die Genauigkeiten liegen jedoch unter der NIR, wodurch die Risikovariablen kein Zugewinn für die Klassifikationsgüte bieten.

Tabelle 2: Modellgüte für die Vorhersage dichotomer Inanspruchnahmevariablen

Inanspruchnahme	Genauigkeit	[95% CI]	NIR	P[Genauigkeit > NIR]
Zusammenarbeit	84,04 %	[75,05 %; 90,78 %]	86,17 %	0,77
Weitervermittlung	63,32 %	[60,55 %; 67,23 %]	64,43 %	0,52
Beendigung	83,89 %	[77,69 %; 88,94 %]	85%	0,71

Anmerkung. NIR = No-Information-Rate

Für die Vorhersagegüte der kontinuierlichen Inanspruchnahmevariablen wird auf das Verhältnis des RMSE zur Standardabweichung des jeweiligen Kriteriums zurückgegriffen (s. Tab. 3, folgende Seite). Zusammenfassend bieten psychosoziale Risikoscreenings einen kleinen Zugewinn für die Vorhersage der Nutzungsintensität (Anzahl Hausbesuche, Netzwerkkontakte) und der Inhalte (Sozialberatung, Eltern-Kind-Interaktion). Angaben zum Inhalt „medizinische Versorgung" lassen sich durch die psychosozialen Risikovariablen nicht ausreichend gut vorhersagen.

3.3 Die Bedeutsamkeit von einzelnen Risikovariablen

In Bezug auf die Inanspruchnahmevariablen, für die Modelle mit einer ausreichenden Vorhersagegüte gefunden wurden, werden im Folgenden die jeweils bedeutsamen Risikovariablen (Prädiktoren) aufgeführt. Abbildung 2 zeigt die relative Wichtigkeit (Importance) der sieben bedeutsamsten Risikovariablen für die Vorhersage des jeweiligen Kriteriums. Die relative Wichtigkeit wurde durch den relativen Einfluss des Prädiktors

auf die Verlustfunktion der GBM berechnet, das heißt, wieviel schlechter die Vorhersage ohne den einzelnen Prädiktor wird. Der wichtigste Prädiktor wird mit dem Wert 100 gekennzeichnet, die Wichtigkeit der anderen Prädiktoren wird im Verhältnis dazu dargestellt. Fünf der sieben wichtigsten Prädiktoren sind die numerischen Gesamteinschätzungen der HBS. Die insgesamt wichtigsten Variablen über alle Modelle hinweg sind die materielle (finanzielle) Belastung und die Belastung des Kindes. Unter den 124 kategorialen Risikofaktoren haben nur der fehlende Kontakt zum Kindsvater und das Vorhandensein von Schulden eine wichtige Bedeutung.

Tabelle 3: Modellgüte metrischer Outcomes

Inanspruchnahme		RMSE	SD	Adj. R^2
Inhalt	Eltern-Kind-Interaktion	0,27	0,30	18,40 %
	Sozialberatung	0,28	0,31	17,48 %
	Medizinische Versorgung	0,23	0,24	7,32 %
Anzahl	Netzwerkkontakte	1,14	1,23	12,37 %
	Hausbesuche	12,95	14,34	14,38 %

Anmerkung. RMSE = Root Mean Squared Error

Zur inhaltlichen Interpretation werden im Online-Material Grafiken des partialisierten Einfluss der wichtigsten Risikovariablen dargestellt. Dieser gibt an, bei welchen Ausprägungen der Prädiktoren welche Werte auf den Inanspruchnahmevariablen zu erwarten sind. Die auffälligsten Trends werden im Diskussionsteil beschrieben.

4 Diskussion

In dieser Studie wurde die Nützlichkeit von psychosozialen Risikoscreenings für die Vorhersage der Inanspruchnahme von Angeboten der Frühen Hilfen untersucht. Dazu wurden in einer Stichprobe von Teilnehmern eines Hausbesuchsprogramms die Güte der Vorhersage von Nutzungsintensität, Interventionsinhalten, Beendigungsgründen, Weitervermittlungen und der Helferbeziehung durch einzelne Risikovariablen und globale Risikoeinschätzungen beurteilt. Zusammenfassend lässt sich eine leicht verbesserte Vorhersage der Nutzungsintensität und der Interventionsinhalte feststellen. In der Vorhersage waren globale Einschätzungen auf numerischen Risikoskalen wichtiger als einzelne kategoriale Risikovariablen.

4.1 Sind psychosoziale Risikoscreenings zur Vorhersage der Inanspruchnahme geeignet?

Für vier der acht untersuchten Inanspruchnahmevariablen lässt sich eine leicht verbesserte Prognosegüte zeigen. Dabei waren die Nutzungsintensität (Zahl der Hausbesuche,

Zahl der Netzwerkkontakte) sowie die Interventionsinhalte (Eltern-Kind-Interaktion, Sozialberatung) ausreichend gut vorherzusagen. Der Interventionsinhalt medizinische Versorgung, die Weitervermittlung an andere Institutionen, Beendigungsgründe sowie die Zusammenarbeit mit der Familie waren nicht ausreichend gut vorherzusagen.

Zur eindeutigen Beurteilung der Vorhersagegüte liegen keine Studien mit einem vergleichbaren methodischen Vorgehen vor. Im Vergleich zu Studien, die in Regressionsmodellen auf vorausgewählte psychosoziale Risikofaktoren zurückgriffen (Daro et al., 2003; Goyal et al., 2016), konnte in dieser Untersuchung ein bedeutend höherer Anteil der Varianz der Inanspruchnahmevariablen erklärt werden. Die Ausnahme bildet eine Untersuchung von Brand und Jungmann (2014), die durch eine Auswahl von neun psychosozialen Risikofaktoren eine bessere Vorhersage des Interventionsabbruchs erreichten, die sie durch das Heranziehen von Prozessvariablen (z. B. elterliche Kooperation) noch verbesserten. Trotz der vergleichsweise besseren Vorhersage im Verhältnis zu den meisten anderen Studien ist die statistische Vorhersagegüte hier als niedrig einzustufen. Neben dem psychosozialen Risiko sollten dementsprechend noch andere Variablen zur Vorhersage der Inanspruchnahme herangezogen werden.

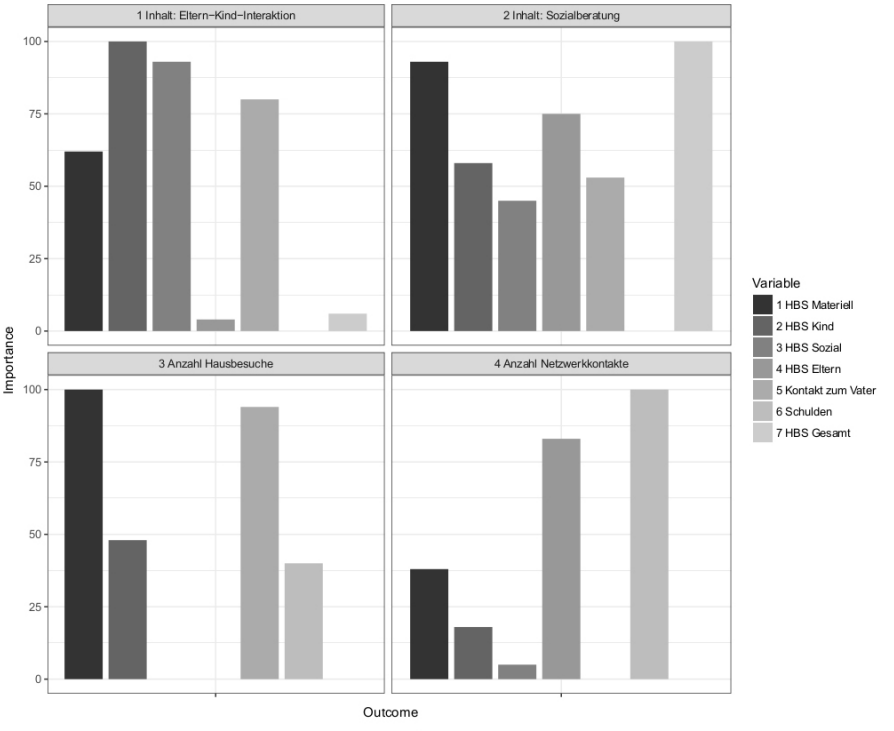
Die inhaltliche Relevanz der Vorhersagegüte lässt sich anhand der Abbildungen 3 bis 9 im Online-Material beurteilen. Daraus wird deutlich, dass einzelne Risikovariablen nur einen geringen Zugewinn in der Vorhersage der Nutzungsintensität bieten. So erklärt die Gesamteinschätzung der materiellen Belastung beispielsweise den Unterschied von plus/minus zwei Hausbesuchen (Abb. 6). Dieser Unterschied fällt im Einzelfall kaum ins Gewicht. Durch die Kombination der wichtigsten psychosozialen Risikofaktoren lässt sich deren individueller Einfluss jedoch aufsummieren. Somit lassen sich beim Vorliegen bestimmter Risikokombinationen Vorhersagen über praxisrelevante Unterschiede treffen. Die Bedeutung der wichtigsten Risikovariablen wird im nächsten Abschnitt erläutert.

4.2 Welche Risikovariablen sagen die Inanspruchnahme gut vorher?

In Abbildung 2 sind die wichtigsten Risikovariablen getrennt nach Bereich der Inanspruchnahme aufgeführt. Der inhaltliche Fokus auf die *Eltern-Kind-Interaktion*, wird am besten durch die Gesamtbelastung des Kindes und die soziale Belastung vorhergesagt. In die Gesamtbelastung des Kindes fließen sowohl medizinische Probleme als auch Verhaltensauffälligkeiten ein. Damit ist der Fokus auf die Eltern-Kind-Interaktion mehr an der Symptomatik des Kindes als an der Symptomatik der Eltern orientiert. Auffallend ist der niedrige Schwellenwert, bei dem schon eine geringe Belastung des Kindes zu einer stärkeren inhaltlichen Schwerpunktsetzung führt (Abb. 3). Dies könnte in den Grundberufen der Fachkräfte begründet sein, durch die sie stärker für die kindlichen Belastungen sensibilisiert wurden (Cierpka, Benz, Doege, Rudolf, 2013). Zudem könnte die Belastung des Kindes, z. B. beim exzessiven Schreien, auch in den Hausbesuchen prominenter in den Vordergrund treten (vgl. Pauli-Pott, Becker, Mertesacker, Beckmann, 2000), sodass stärker auf die Reaktion auf kindliche Signale fokussiert wird. Bei einer hohen sozialen Belastung, das heißt keiner sozialer Unter-

stützung oder einem dissozialen Umfeld, sinkt der Fokus auf die Eltern-Kind-Interaktion deutlich (Abb. 4). In diesem Fall könnten in der Intervention mehr praktische Anliegen in den Vordergrund treten (Kitzman et al., 1997).

Abbildung 2: Relative Wichtigkeit der Risikovariablen für die Inanspruchnahmebereiche



Anmerkungen. Prädiktoren sind nach absoluter Wichtigkeit über alle Modelle hinweg geordnet. HBS = Heidelberger Belastungsskala (globale Belastungsbeurteilung)

Im inhaltlichen Schwerpunkt *Sozialberatung* tritt der Prädiktor der materiellen Belastung in den Vordergrund. Die inhaltliche Beschäftigung mit Sozialberatung nimmt im Bereich mittlerer Belastung zu und bei einer sehr hohen Belastung wieder ab (Abb. 5). Der Großteil der aufsuchenden Helferinnen kam aus dem Gesundheitsbereich und erhielt eine kurze Weiterbildung in sozialrechtlichen Fragen (Cierpka et al., 2013). Damit könnten die Fachkräfte im mittleren Belastungsbereich grundlegende Fragen bezüglich finanzieller Hilfe besprochen haben. Im hohen Belastungsbereich, das heißt bei großer Armut und Wohnungsenge, griffen die Fachkräfte möglicherweise auf externe Beratung zurück, sodass wieder andere Inhaltsbereiche

besprochen werden konnten. Für die psychosoziale Gesamtbelastung ist schon ab einer mittleren Belastung ein deutlicher Anstieg im Schwerpunkt Sozialberatung zu verzeichnen. Möglicherweise schlagen sich finanzielle Probleme besonders deutlich in der familiären Stressbelastung nieder, sodass hier die Gewährleistung der akuten Versorgung in den Vordergrund getreten ist (Yates, Obradović, Egeland, 2010).

Die *Zahl der Hausbesuche* lässt sich vor allem durch die materielle Belastung und den Kontakt zum Kindsvater vorhersagen. Eine mittlere und hohe materielle Belastung hängt dabei mit einer höheren Gesamtzahl an Hausbesuchen zusammen (Abb. 6). Die Belastung der Kinder hat im Vergleich dazu einen geringen Einfluss, die elterliche und familiäre Belastung hat keinen Einfluss auf die Interventionsdosis. Dies widerspricht internationalen Befunden (Daro et al., 2003; Goyal et al., 2016) und könnte für eine notwendige anfängliche Stabilisierungsphase bei finanziellen Problemen sprechen, bevor die Eltern-Kind-Interaktion thematisiert werden kann. Bei bestehendem Kontakt der Mütter zum Kindsvater nimmt die Zahl der Hausbesuche ab (Abb. 7). Da Paarkonflikte im Modell keinen Einfluss hatten, scheint es sich beim Kontakt zum Vater um eine stabilisierende Ressource zu handeln, die den Unterstützungsbedarf verringert.

Für die *Anzahl der Kontakte zu anderen Institutionen* im Netzwerk Frühe Hilfen ist das Vorliegen von Schulden als isolierter Risikofaktor am bedeutsamsten. Haben die Familien Schulden, werden marginal mehr Institutionen einbezogen (Abb. 8). Ähnlich wie beim Schwerpunkt Sozialberatung könnte dies durch die Grundberufe der aufsuchenden Helferinnen zu erklären sein. Eine sehr hohe elterliche und familiäre Belastung, die durch schwere Familienkonflikte, Gewalt oder psychische Störungen gekennzeichnet ist, hängt mit einer deutlichen Steigerung der Netzwerkkontakte zusammen (Abb. 9). Hier könnte der Aspekt Kinderschutz greifen, der ebenfalls integraler Bestandteil der Frühen Hilfen ist (Cierpka u. Evers, 2015).

In der übergreifenden Bewertung der Risikofaktoren ist auffällig, dass die globalen HBS-Skalen aufgrund ihrer Eigenschaft als Gesamtrisikoeinschätzung für die Voraussage der Inanspruchnahme wichtiger zu sein scheinen als einzelne Risikofaktoren. Analogien finden sich in der Bewertung des Misshandlungsrisikos, das durch kumulative Risikomodelle besser erklärt werden kann als durch Einzelrisiken (Begle, Dumas, Hanson, 2010). Im Gegensatz zur Aufsummierung von Einzelrisiken ist die globale Bewertung auf der HBS jedoch sparsamer. Allerdings könnte die vorherige Einschätzung der Einzelrisiken in der HBS die Güte der globalen HBS-Ratings positiv beeinflussen. Ein weiterer psychometrischer Vorteil der Gesamtskalen ist die feinere Abstufung von 0 bis 100, die von den aufsuchenden Helferinnen in der gesamte Breite der Skala genutzt wurde.

4.3 Welche Bereiche der Inanspruchnahme werden nicht gut vorhergesagt?

Weitervermittlungen an andere Institutionen ließen sich nicht gut vorhersagen. Das spricht dafür, dass Netzwerkstrukturen allgemein für Weiterempfehlungen genutzt wurden und psychosoziale Risiken eher einen Einfluss auf die Intensität der Netz-

werkkontakte hatten. In einer Studie zur Rolle der Jugendhilfe und der Frühen Hilfen in Präventionsketten haben Evers und Cierpka (2015) bereits dargestellt, dass sich verschieden belastete Subgruppen nicht im Anteil der Weitervermittlungen unterschieden, sondern in der Art der Institutionen, an die weiterverwiesen wurde.

Weder die *Güte der Zusammenarbeit noch irreguläre Beendigungen*, wie z. B. Interventionsabbrüche, konnten durch die psychosoziale Belastungen ausreichend vorhergesagt werden. Ähnliche Ergebnisse finden sich auch im Programm „Pro Kind", bei dem Prozessvariablen den Interventionsabbruch besser vorhersagten als psychosoziale Risiken (Brand u. Jungmann, 2014). In der psychotherapeutischen Versorgung lassen sich ähnliche Befunde zum untergeordneten Einfluss von sozioökonomischem Status auf die therapeutische Beziehung und Klientenzufriedenheit finden (Kapp et al., 2017).

4.4 Limitationen und Ausblick

Die vorliegende Studie beschränkt sich auf ein Programm der Frühen Hilfen. Daher sollten, trotz der Kreuzvalidierung innerhalb des vorliegenden Datensatzes die Ergebnisse auch in anderen Settings und Regionen auf die Replizierbarkeit überprüft werden. Eine mögliche Einschränkung bildet auch die weite Auffächerung der HBS mit 124 einzelnen Risikofaktoren. Durch die starke Differenzierung könnte die Prävalenz bestimmter Risiken unterschätzt worden sein, wodurch sie im Vergleich zu den Gesamtskalen an Bedeutung verlieren würden. Eine explorative Zusammenfassung einzelner Risiken nach Belastungsbereich brachte in Voranalysen jedoch keine Verbesserung der Modellgüte.

Aussagen zur Indikationsstellung sind dadurch eingeschränkt, dass die Daten ausschließlich auf der Inanspruchnahmepopulation des Programms KfdN beruhen. Die hier verwendete statistische Methode zeigt sich jedoch als geeignet, um in zukünftigen Untersuchungen die Studienpopulation, Prädiktoren und vorhergesagte Variablen zu erweitern. Zur besseren Generalisierung auf die breite Population von Familien in den Frühen Hilfen wäre das Hinzuziehen von Daten mehrerer Programme, Träger und Versorgungsgebiete sinnvoll. Dazu müsste eine einheitliche Datengrundlage an Kernvariablen für die Erfassung der Inanspruchnahme erarbeitet werden. Ebenso sollte die Erhebung um Familien erweitert werden, die trotz anfänglicher Indikation kein Angebot der Frühen Hilfen in Anspruch nehmen. Eine Nachbefragung von Teilnehmern, die das Programm abbrechen, wäre zur besseren Differenzierung der Abbruchgründe zu empfehlen. Mithilfe der verwendeten statistischen Methode ließen sich die Prädiktoren um weitere psychometrische Instrumente ergänzen (z. B. zur Psychopathologie, Selbstwirksamkeit, Kontrollüberzeugung, Bindung, familiäres Funktionsniveau).

In zukünftigen Studien könnte das Kriterium um die Prognose einer differenziellen Wirkung der Intervention ergänzt werden. Wie im Bereich der Inanspruchnahmeforschung liegen hier bisher nur Studien zu isolierten Risikofaktoren als Prädiktoren der Wirksamkeit vor (Caldera et al., 2007; Miller, Farkas, Duncan, 2016). Da in dieser Studie globale Risikoeinschätzungen von höherem prädiktiven Wert waren als kategoriale Ein-

zelrisiken, könnten zukünftige Studien auch in der Einschätzung des Misshandlungsrisikos die prädiktive Validität von globalen Ratings mit Einzelratings vergleichen.

4.5 Zusammenfassung und Fazit

In der aktuellen Studie konnte die Inanspruchnahme von Frühen Hilfen auf der Basis von psychosozialen Risikofaktoren besser vorhergesagt werden als in vorherigen Studien, die auf einzelnen Risikovariablen beruhten. Die Vorhersagegüte ist dennoch nicht zufriedenstellend, was für die Präsenz anderer bedeutsamer Einflüsse auf die Inanspruchnahme spricht.

Inhaltlich relevante Vorhersagen lassen sich zu Interventionsinhalten und Intensität treffen. Diese hingen hier besonders von globalen Bewertungen der Risikobereiche ab und weniger von Einzelrisiken. Dabei stechen vor allem finanzielle Belastungen heraus. Stärkere Belastungen scheinen dabei im Allgemeinen mit einer höheren Interventionsdosis und einer inhaltlichen Fokusverschiebung einherzugehen. Einzig die Belastungseinschätzung des Kindes bietet einen sehr sensitiven Indikator für die Fokussierung auf die Eltern-Kind-Interaktion.

Fazit für die Praxis

Auf Grundlage der hier gefundenen Ergebnisse können psychosozialen Risikoscreenings als Einzelinstrument zur differenziellen Indikationsstellung nicht empfohlen werden. Dagegen spricht vor allem das ungünstige Verhältnis von prognostischer Sicherheit und Aufwand. Sofern Risikoscreenings jedoch ebenfalls zur Gefährdungseinschätzung eingesetzt werden, können sie wichtige Zusatzinformationen zur voraussichtlichen Inanspruchnahme liefern.

Wird eine hohe finanzielle Gesamtbelastung oder ein hohes Gesamtrisiko festgestellt, könnte bereits vor Interventionsbeginn oder im Tandemmodell (Brand u. Jungmann, 2012) eine fokussierte Sozialberatung hinzugezogen werden. Damit bestünde die Möglichkeit, sich stärker auf originäre Programminhalte zu fokussieren und aufsuchende Helferinnen zu entlasten. Insbesondere bei einer hohen finanziellen Gesamtbelastung, elterlicher Belastung, dem Vorliegen von Schulden und Abwesenheit des Kindsvaters ist zudem ein höherer Interventions- und Kooperationsaufwand zu erwarten. Diese Faktoren könnten besonders in der Planung des Caseloads beachtet werden. Insgesamt könnten psychosoziale Risikoeinschätzungen in Kombination mit klinischen Urteilen und einem Monitoring des Interventionsprozesses einen umfassenden Blick auf die Inanspruchnahme der Frühen Hilfen bieten.

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

Literatur

Begle, A. M., Dumas, J. E., Hanson, R. F. (2010). Predicting Child Abuse Potential: An Empirical Investigation of Two Theoretical Frameworks. Journal of Clinical Child & Adolescent Psychology, 39, 208-219.

Bender, D., Lösel, F. (2014). Risikofaktoren, Schutzfaktoren und Resilienz bei Misshandlung und Vernachlässigung. In U. T. Egle, P. Joraschky, A. Lampe, I. Seiffge-Krenke, M. Cierpka (Hrsg.), Sexueller Missbrauch, Misshandlung und Vernachlässigung (4. Aufl., S. 77-103). Stuttgart: Schattauer.

Brand, T., Jungmann, T. (2010). Abschlussbericht der Implementationsforschung zum Modellprojekt „Pro Kind". Hannover: Criminological Research Institute of Lower Saxony.

Brand, T., Jungmann, T. (2012). Implementation differences of two staffing models in the German home visiting program "Pro Kind". Journal of Community Psychology, 40, 891-905.

Brand, T., Jungmann, T. (2014). Participant characteristics and process variables predict attrition from a home-based early intervention program. Early Childhood Research Quarterly, 29, 155-167.

Caldera, D., Burrell, L., Rodriguez, K., Crowne, S. S., Rohde, C., Duggan, A. (2007). Impact of a statewide home visiting program on parenting and on child health and development. Child Abuse & Neglect, 31, 829-852.

Cierpka, M. (2009). Keiner fällt durchs Netz. Wie hoch belastete Familien unterstützt werden können. Familiendynamik, 34, 156-167.

Cierpka, M., Benz, M., Doege, D., Rudolf, M. (2013). Frühe Hilfen – Keiner fällt durchs Netz: Bilanzbericht Projektlaufzeit 2007 – 2011. Saarbrücken.

Cierpka, M., Evers, O. (2015). Implementation and efficacy of early-childhood interventions in German-speaking countries. Mental Health & Prevention, 3, 67-68.

Cierpka, M., Gregor, A., Frey, B. (2004). Das Baby verstehen. Heidelberg: Focus-Familie gGmbH.

Daro, D., McCurdy, K., Falconnier, L., Stojanovic, D. (2003). Sustaining new parents in home visitation services: key participant and program factors. Child Abuse & Neglect, 27, 1101-1125.

Duggan, A., Windham, A., McFarlane, E., Fuddy, L., Mph, L., Rohde, C., Buchbinder, S., Sia, C. (2000). Hawaii's Healthy Start Program of Home Visiting for At-Risk Families: Evaluation of Family Identification, Family Engagement, and Service Delivery. Pediatrics, 105(Supplement 2), 250.

Evers, O., Cierpka, M. (2015). Pathways in prevention–subgroups in an early preventive intervention program and their engagement with the child welfare service. Mental Health & Prevention, 3, 117-128.

Fraser, J. A., Armstrong, K. L., Morris, J. P., Dadds, M. R. (2000). Home visiting intervention for vulnerable families with newborns: follow-up results of a randomized controlled trial. Child Abuse & Neglect, 24, 1399-1429.

Girvin, H., DePanfilis, D., Daining, C. (2007). Predicting Program Completion Among Families Enrolled in a Child Neglect Preventive Intervention. Research on Social Work Practice, 17, 674-685.

Goyal, N. K., Folger, A. T., Hall, E. S., Teeters, A., Van Ginkel, J. B., Ammerman, R. T. (2016). Multilevel assessment of prenatal engagement in home visiting. Journal of Epidemiology and Community Health, 70, 888.

Hahn, M., Sandner, E. (2013). Kompetenzprofil Familienhebammen. Köln.

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

Kapp, C., Perlini, T., Jeanneret, T., Stéphan, P., Rojas-Urrego, A., Macias, M., Halfon, O., Holzer, L., Urben, S. (2017). Identifying the determinants of perceived quality in outpatient child and adolescent mental health services from the perspectives of parents and patients. European Child & Adolescent Psychiatry, 26, 1269-1277.

Kindler, H. (2009). Wie könnte ein Risikoinventar für Frühe Hilfen aussehen. In T. Meysen, L. Schönecker, H. Kindler (Hrsg.), Frühe Hilfen im Kinderschutz. Rechtliche Rahmenbedingungen und Risikodiagnostik in der Kooperation von Gesundheits- und Jugendhilfe (S. 171-261). Weinheim: Juventa.

Kindler, H. (2010). Risikoscreening als systematischer Zugang zu Frühen Hilfen. Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz, 53, 1073-1079.

Kitzman, H. J., Cole, R., Yoos, H. L., Olds, D. (1997). Challenges experienced by home visitors: A qualitative study of program implementation. Journal of Community Psychology, 25, 95-109.

Künster, A. K., Knorr, C., Fegert, J. M., Ziegenhain, U. (2010). Soziale Netzwerkanalyse interdisziplinärer Kooperation und Vernetzung in den Frühen Hilfen. Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz, 53, 1134-1142.

McCurdy, K., Daro, D., Anisfeld, E., Katzev, A., Keim, A., LeCroy, C., McAfee, C., Nelson, C., Falconnier, L., McGuigan, W. M., Park, J. K., Sandy, J., Winje, C. (2006). Understanding maternal intentions to engage in home visiting programs. Children and Youth Services Review, 28, 1195-1212.

McGuigan, W. M., Katzev, A. R., Pratt, C. C. (2003). Multi-level determinants of retention in a home-visiting child abuse prevention program. Child Abuse & Neglect, 27, 363-380.

Miller, E. B., Farkas, G., Duncan, G. J. (2016). Does Head Start differentially benefit children with risks targeted by the program's service model? Early Childhood Research Quarterly, 34, 1-12.

Moore, P. D., Bay, R. C., Balcazar, H., Coonrod, D. V., Brady, J., Russ, R. (2005). Use of Home Visit and Developmental Clinic Services by High Risk Mexican-American and White Non-Hispanic Infants. Maternal and Child Health Journal, 9, 35-47.

NZFH (Hrsg.) (2014). Bundesinitiative Frühe Hilfen: Zwischenbericht 2014 Köln.

Olds, D. L., Korfmacher, J. (1998). Maternal psychological characteristics as influences on home visitation contact. Journal of Community Psychology, 26, 23-36.

Pauli-Pott, U., Becker, K., Mertesacker, T., Beckmann, D. (2000). Infants with "Colic" – mothers' perspectives on the crying problem. Journal of Psychosomatic Research, 48, 125-132.

Sidor, A., Eickhorst, A., Stasch, M., Cierpka, M. (2012). Einschätzung der Risikobelastung in Familien im Rahmen von Frühen Hilfen: Die Heidelberger Belastungsskala (HBS) und ihre Gütekriterien. Praxis der Kinderpsychologie und Kinderpsychiatrie, 61, 766-780.

Suess, G. J., Bohlen, U., Mali, A., Frumentia Maier, M. (2010). Erste Ergebnisse zur Wirksamkeit Früher Hilfen aus dem STEEP-Praxisforschungsprojekt „WiEge"1. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 53, 1143-1149.

van der Put, C. E., Assink, M., Boekhout van Solinge, N. F. (2017). Predicting child maltreatment: A meta-analysis of the predictive validity of risk assessment instruments. Child Abuse & Neglect, 73, 71-88.

Wiegand-Grefe, S., Halverscheid, S., Plass, A. (2011). Kinder und ihre psychisch kranken Eltern: familienorientierte Prävention-der CHIMPs-Beratungsansatz. Göttingen: Hogrefe.

Yates, T. M., Obradović, J., Egeland, B. (2010). Transactional relations across contextual strain, parenting quality, and early childhood regulation and adaptation in a high-risk sample. Development and Psychopathology, 22, 539-555.

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

Ziegenhain, U. (2007). Förderung der Beziehungs- und Erziehungskompetenzen bei jugendlichen Müttern. Praxis der Kinderpsychologie und Kinderpsychiatrie, 56, 660-675.

Zusatzmaterial zu diesem Beitrag finden Sie zum Herunterladen auf der Detailseite von Jahrgang 67 der „Praxis der Kinderpsychologie und Kinderpsychiatrie", Heft 67,5, unter www.vandenhoeck-ruprecht-verlage.com

Korrespondenzanschrift: Oliver Evers, Universitätsklinikum Heidelberg, Institut für Psychosoziale Prävention, Bergheimer Str. 54, 69115 Heidelberg;
E-Mail: oliver.evers@med.uni-heidelberg.de

Oliver Evers und *Paul Schröder*, Universitätsklinikum Heidelberg, Institut für Psychosoziale Prävention, Heidelberg

Prax. Kinderpsychol. Kinderpsychiat. 67: 462 – 480 (2018), ISSN: 0032-7034 (print), 2196-8225 (online)

Study 2

II. Georg, A.K., **Schröder-Pfeifer, P**., Cierpka, M. & Taubner, S. (Under review). Parenting stress in the face of early regulatory disorders in infancy – what matters most? *Journal of Developmental and Behavioural Pediatrics*. IF: 2.056

Declaration of author contributions: Georg, A.K.: Conceptualization, Methodology, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Schröder-Pfeifer, P.:** Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Formal analysis, Visualization **Cierpka, M.:** Conceptualization, Supervision, Project Administration, Funding Acquisition **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

Parenting stress in the face of early regulatory disorders in infancy: What matters most?

Georg, A.K., Cierpka, M., Taubner, S. & Schröder-Pfeifer, P.,

Abstract

Objective: Early regulatory disorders (ERD) in infancy are typically associated with high parenting stress. Given the range of factors that may contribute to parenting stress, clinicians aiming to help burdened families must administer extensive assessments of infant symptoms, current parental psychological symptoms, as well as a host of other risk and protective factors. The aim of this study was to identify key predictors of parenting stress in a sample of $N = 135$ mothers from infants diagnosed with ERD.

Methods: We used machine learning algorithms to analyze the data. Parenting stress was assessed with the Parenting Stress Index. The multivariate dataset consisted of 464 variables covering mother-reported psychological distress, maternal self-efficacy, parental reflective functioning, socio-demographics, each parents' history of illness, recent significant life events, former miscarriage/abortion, pregnancy, obstetric history, infants' medical history, development, and social environment. Behavioral diaries assessed infants' regulatory symptoms and parental co-regulative behavior. A clinical interview was utilized to diagnose ERD and to assess regulatory symptoms, quality of parent-infant relationship, organic/biological and psychosocial risks, and social-emotional functioning.

Results: The final prediction model identified 11 important variables summing up to maternal self-efficacy, psychological distress (especially depression and anger-hostility), infant regulatory symptoms, and age-appropriate physical development. The RMSE (i.e., prediction accuracy) of the final model applied to the test set was 21.72 ($R^2 = 0.58$).

Conclusions: With these predictors identified, clinicians can more efficiently assess a mother's parenting stress related to ERD with mainly sleeping disorders in a low-risk sample.

Key words: early regulatory disorders, machine learning algorithms, parenting stress, parental self-efficacy, family diagnostic

Introduction

Early regulatory disorders (ERD), which include sensory, sleeping, crying or feeding disorders, are found in approximately 10.9% of infants/toddlers and are among the most prevalent diagnoses in children under the age of four.[1] The disorders have been repeatedly found to be associated with high parenting stress (PS) and parental burden.[2, 3] Research has focused on the effects of excessive crying and infant colic: E.g., compared to control groups, mothers reported higher negative affect in response to the cries[4] and felt more sad and aroused by the cries.[5] According to the developmental systems model of ERD, the parental stress response to infants' regulation problems may contribute to a vicious circle that perpetuates parental burden, impairs parental self-efficacy, and leads to the manifestation or perpetuation of ERD.[6]

While the disorders likely have far reaching consequences for a child [7], they do not necessarily have such effects: Smarius et al. found that the maternal burden of infant care partially mediated the association between ERD and later mood and behavioral problems in childhood.[8] In addition, mothers' PS predicted the persistence of regulation problems.[9] These studies suggest that reducing PS may be an effective objective in treating ERD.

Predictors of parenting stress in early regulatory disorders

While the adverse effects of ERD on parents have been established, the specific risk and protective factors associated with PS when raising infants with ERD have yet to be explored. A set of risk factors for parents' propensity to experience PS in the context of ERD have been proposed:[6, 10] high prenatal maternal stress or lifetime depressive or anxiety disorders have been found to predict ERD,[11, 12] and sociodemographic risk factors, such as low social support or low maternal education have been shown to be related to ERD[9, 11, 13] and may negatively affect PS.[14, 15]

PS has also been linked to miscarriages or abortions,[15] which have been found to be more prevalent in a clinical ERD sample.[6] Peripartum risk-factors, like complicated pregnancy or birth which are more frequent in ERD samples,[6, 16] may affect parents' perception of infants'

distress and thus their propensity to experience PS. Other infant diagnostic characteristics, such as the presence of an organic condition or difficult temperament, were related to ERD[6, 13] and may increase PS.[17] Maternal self-efficacy was a protective factor for reporting ERD[11, 18] and may protect against high PS. Similarly, parental reflective functioning may be a protective factor for high PS related to ERD.[19]

While this literature provides valuable data on distinct predictors and correlates of ERD, the extent to which these variables moderate PS in the context of ERD have rarely been investigated. Furthermore, few of the studies included samples of infants with ERD beyond excessive crying. An additional limitation of the literature is the inclusion of a small number of variables, despite multiple and interrelated factors within a family system. Thus, in order to find key ports of entry, clinicians wishing to help families who experience ERD must assess a vast number of possible variables. Thus, the goal of this study was to identify key variables associated with PS in ERD. To this end, machine learning (ML) algorithms were applied.

Machine Learning approaches in clinical psychology and psychiatry

ML approaches for clinical psychology and psychiatry perform statistical functions on multidimensional data sets to make generalizable predictions about individuals. That is, ML provides estimates on how well the obtained results of a prediction model may be generalized to an individual, which in our study is a future parent. In the field of child psychiatry, ML may prove especially useful for the incorporation of data from different sources and developmental factors.[20] The algorithms utilized by ML can integrate large sets of correlated variables, are insensitive to outliers, and assume no distribution in the outcome or underlying data mechanism.[21] In addition, prediction models can include data on single item level, which results in an item selection that can be utilized to shorten clinical and diagnostic batteries. All of these features are especially useful for the aim of this study.

The present study

We employed ML in an exploratory search for variables best predicting a mother's PS related to ERD. Predictor variables were empirically and theoretically derived[6, 10] and covered risk and protective factors, as well as correlates that have been identified for ERD or PS. We included a multivariate dataset by utilizing multiple measures: mother-reported general psychological distress, maternal self-efficacy, parental reflective functioning, socio-demographic variables, each parents' history of illness, recent significant life events, former miscarriage/abortion, pregnancy, obstetric history, infants' medical history, development, and social environment. Behavioral diaries were used to assess infants' regulatory symptoms, and extent of parental co-regulative behavior. A structured clinical interview was utilized to diagnose ERD and to assess regulatory symptoms, quality of the parent-infant relationship, organic/biological and psychosocial risks, and social-emotional functioning.

In addition to global scores obtained from the instruments, we analyzed all items gathered in our dataset on single item and subscale levels in an effort to maximize specificity of the predictors.

Method

Data was collected from February 2014 to May 2017 in the department for [blinded] and stemmed from a RCT on the effectiveness of brief parent-infant psychotherapy for ERD, where data collection was still ongoing by the time of this study. We used data gathered pre-treatment at one time point.

The approval for research in this sample was obtained from the Ethical Committee of [blinded] (approved in November 2013).

Participants

Families were referred from pediatric practices for the purpose of study participation if parents reported significant crying, sleeping or feeding difficulties. Some families self-referred in response to public advertisement, websites, and flyers/posters distributed in gynecological, pediatric and osteopathic practices, parent-infant groups, and crèches.

Inclusion criteria required the infant to be between 4 and 15 months old, born at full term (>37 weeks of gestation), and to meet diagnostic criteria for sleeping disorders, feeding disorders, or regulation disorders of sensory processing according to DC:0-3 R[22] or for persistent excessive crying, sleeping and feeding disorder, according to the guidelines recommended by the German Society of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy (AWMF-guidelines; AWMF No. 028/028).[23] Pregnancy needed to be singleton and primary caregivers needed to speak German.

Participants were excluded when infants had a medical diagnosis that better explained the regulatory problems, a tentative diagnosis of fetal alcohol syndrome, or a diagnosed disability or developmental disorder. A very high symptom severity of the primary caregiver (Symptom-Check-List-90R-S, Global Severity Index of $T>70$)[24] also led to exclusion, as a current mental illness was considered to be a contra-indication for brief interventions.

A total of 165 primary caregivers expressed their interest in study participation and underwent screening for eligibility via telephone. Parents were informed about the study and invited for participation if they consented. Of these, 24 cancelled or did not show up. Six families fulfilled exclusion criteria and thus were excluded. The primary caretaker was asked to participate, which in all cases was the mother. The final sample consisted of $N = 135$ mother-infant dyads.

Procedure and assessments

Self-report measures and behavioral diaries were mailed to mothers following the phone screen. Clinical diagnostics were led by two psychologists. The assessment was conducted with mother and infant and included the clinical interview and video recording of standardized parent-infant interactions. Written informed consent was gathered at the beginning of the session. Clinical ratings were performed immediately after the interview.

The employed measures are described below.

Parenting stress. The Parenting Stress Index[25] assesses self-reported PS with 48 items. Items are rated on a five-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Higher scores indicate higher PS. Items are summed up into one global score. Cronbach's α was excellent in this study (.94).

Psychological distress. The Symptom-Checklist (Symptom-Checklist-90R-S, SCL)[24] assesses self-reported psychological symptoms. The 90 items are rated on a five-point Likert scale from 0 (*not at all*) to 4 (*extremely*) with higher scores indicating higher distress. Items add up to 10 subscales (somatization, obsession-compulsion, interpersonal sensitivity, depression, anxiety, anger-hostility, phobic anxiety, paranoid ideation, psychoticism, and additional clinical symptoms), a sum score, and the global severity index (GSI). Cronbach's α was between .56 (psychoticism) and .84 (obsession-compulsion) and was excellent for the sum score (.96).

Maternal self-efficacy. The Maternal Self-Efficacy Scale (MSES)[26] assesses self-reported behavioral competence in parenting. For this study, back-translation procedures were implemented, and the final version was reviewed by an English native speaker. The 10 Items are rated on a four-point Likert scale from *not good at all* (1) to *very good* (4). Cronbach's α was acceptable (.75).

Parental reflective functioning. The Parental Reflective Functioning Questionnaire (PRFQ)[27] uses 18-items in order to assesses the scales: (1) interest and curiosity in mental states (IC), (2) certainty of mental states (CMS), and (3) prementalizing (PM). Items are rated on a seven-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (7). Cronbach's α was acceptable for CMS (.73), poor for PM (.57), and unacceptable for IC (.47).

Parent-Questionnaire. The Parent-Questionnaire[28] was developed for the assessment of parents and their infants with ERD. Questions refer to sociodemographic information, history of illness, recent significant life events, former miscarriage/abortion, pregnancy, obstetric history, and infant medical history, development, and social environment. Variables are assessed dimensionally and categorically or in open format; no sum scores are provided. For

the analysis, 110 single items were used (see Supplemental Digital Content 1, which lists the items).

Clinical interview. A structured clinical interview was developed to assess axis I (DC:0-3R)[22] on sleep onset disorder, night-waking disorder, feeding disorders, and regulation disorders of sensory processing. Persistent excessive crying syndrome is not mentioned as a clinical category in DC:0-3R and diagnostic criteria are poorly described.[22] Therefore, we additionally utilized the AWMF-guidelines on persistent excessive crying, sleep onset disorder, night-waking disorder, feeding disorders, and pervasive regulatory disorder (AWMF)[23]. The parent-infant relationship global assessment scale (PIR-GAS, DC:0-3R) dimensionally assesses parent-infant relationship from *documented maltreatment* (0-10) to *well adapted* (91-100). Medical conditions of the infant (axis III of DC:0-3R) and psychosocial stressors (axis IV of DC:0-3R) were dimensionally assessed using organic/biological and the psychosocial risk scales.[29] Infants' emotional and social functioning (axis V of DC:0-3R) was rated on the proposed rating scale (DC:0-3R). In sum, 150 variables covering single symptoms, sum-scores of symptoms on the level of diagnosis and axis, as well as a general symptom sum score were used in analysis (see Supplemental Digital Content 1).

Infant regulatory symptoms. The Questionnaire for Crying, Feeding, and Sleeping (QCFS)[30] assesses crying, sleeping and feeding symptoms and parents' dysfunctional co-regulation behavior in response to the symptoms (e.g. "only falls asleep when being carried"). The 53 items constitute the three scales (1) fussing/crying and sleeping, (2) feeding, (3) dysfunctional co-regulation, and a global score. Higher scores indicate more symptoms, parental burden, and dysfunctional co-regulation. Frequency questions are rated on a four-point Likert scale from *never/rarely* (1) to *always/every day* (4). Parents' perceived difficulty are rated from *not at all* (1) to *a lot* (4). Cronbach's α good for the scales (scale 1 = .82; scale 2 = .76; scale 3 = .84) and the global score (.82).

96-hour behavior diary. The diary of crying, sleeping and feeding behavior[31] is similar to widely used parental diaries and assesses infants' behavior and parents' co-regulation behavior. Frequency and duration of fussing/crying, sleeping/waking, feeding, and parental co-regulation is recorded in 15-minute intervals on four consecutive days. Additional questions refer to the success of parental co-regulative strategies. In sum, 139 variables were used in the analysis (see Supplemental Digital Content 1).

Infant development. The Ages and Stages Questionnaire (ASQ-3)[32] is a series of 21 parent-rated questions on children's developmental performance in communication, gross motor, fine motor, problem solving and personal-social skills, represented on five scales. The 30 items are rated with regards to the child's competence as *yes* (10), *sometimes* (5) or *not yet* (0). We used the German translation of the questionnaires for 4, 6, 8, 9, 10, and 14 months old infants. Internal consistency of the scales was not calculated, due to some small age-dependent subgroups. Other studies have shown it to be poor to excellent.[32]

The PSI, SCL, and the QCFS are valid, reliable measures. For the MSES, PRFQ, and ASQ validity and reliability have only been demonstrated for the original English version.

Statistical Analysis

For the prediction of the PSI, all data provided by questionnaires, behavioral diary, and clinical interview on the level of items, subscales, and global scores were used, resulting in 596 variables. Of these, variables with less than 50% missing values before imputation were used, resulting in a final set of 464 variables. The remaining data contained 5.48% missing values. Imputation was done assuming missing at random after visual inspection of pattern of missingness plots. Multiple imputations by chained equations,[33] using fully conditional specification with 40 iterations, were utilized to produce asymptotically unbiased estimations of the data.

An important difference between ML approaches and more commonly used statistical methods is the absence of *p* values and, furthermore, in-sample model fit as a measure of

"success". In ML, the main statistic of interest is the prediction accuracy which is why there are usually two phases: Training the algorithm and testing the result for generalizability. To this end, data in our study was split into a training set containing 70% of all cases and a test set containing the remaining 30%. All statistical analyses were performed with R version 3.5.2.[34] The R package "caret" version 6.0-76 was used to train the algorithms.

In the first phase, we trained our ML algorithm on the training set in order to select the best performing algorithm. Algorithms were trained using 5-fold cross-validation and 10 repeats. Predicted values of the PSI that the algorithm would assign to the left out fifth were compared with the observed PSI values in that sample. The difference was computed and averaged (in our case, root mean square error, RMSE, as well as mean absolute error MAE were calculated) over all observations. This process was repeated with every fifth and the result was averaged over all iterations. This was then, in turn, repeated 10 times with different splitting points for the data.

In the second phase (i.e., test phase), the algorithm best performing was further tested for generalizability. Prediction accuracy was computed with the hold out test sample of the remaining 30% of cases and by comparing the predicted values with the observed PSI values.

Feature selection. To improve prediction performance, we used a recursive backwards selection, based on importance ranking of random forests, out of the entire set of 464 variables. The result was a set of 11 variables that were deemed to be most informative in terms of PSI and which were used in combination with the algorithm as described below. The FS "rfe" function from the caret package was used to implement this.

Gradient Boosting Machines (GBM). The hyperparameter grid search for the GBM was done by iteratively manipulating the shrinkage coefficient (eta) between 0.01 and 0.2, the interaction depth of each tree (max_depth) between one and six, the number of boosting iterations (nrounds) between one and 1500, while keeping the minimum loss reduction (gamma) fixed at zero and the minimum sum of instance weight (min_infant_weight) fixed at one. The

final values for the model were eta = 0.01, nrounds = 500, max_depth = 1. The gradient boosting model "xgbTree" from the caret package was used.

In an effort to rank the predictors of PS according to their importance, we analyzed the variable importance of the final GBM model. Importance was calculated as the relative influence of the variable on the reduction in the loss function of the GBM model. The most important variable was assigned the value of 100 while the others were scaled accordingly.

In order to assess the marginal effects in which the variables influenced PSI, we looked at partial dependency plots of the most important variables and their interrelation in predicting PSI. Marginal effects were calculated using Friedman's tree traversal method.[35]

In order to evaluate possible interaction effects in the GBM, the procedure described in Lampa et. al. were applied.[36] No significant interactions were found.

Results

Participants

Table 1 gives an overview of the sample characteristics. On average, the parent-infant relationship was rated as perturbed (PIR-GAS, 71-80). The percentage of maternal lifetime mental illness was lower compared to the lifetime prevalence rates in Germany (25.2%).[37] Mothers' average psychological distress (SCL-GSI) was equivalent to a *T*-score of 57, which is approximately > 1 *SD* higher compared to the normative sample.[24] On average, they experienced more PS (PSI) than 88% of the normative sample.[25]

INSERT TABLE 1 HERE

Performance

The model with FS was significantly better than the model without FS (GBM vs GBM with FS: $t(15.3) = 3.4$; $p < .01$). Thus, GBM with FS was utilized for all final results.

The RMSE (i.e., prediction accuracy) of the final model applied to the test set was 21.72, the R^2 was .58 and the MAE was 17.04. Thus, the algorithm on average over- or underestimated the observed PSI score of the participants by 17.04 points or within 10.72% of the observed

PSI range which was 159. The relatively small difference between RMSE and MAE indicates that there were few observations that had larger than average residuals.

Importance of Variables

Figure 1 displays the relative importance of the variables in predicting PSI. Among the most important predictors were maternal self-efficacy (MSES sum score) and two items of the SCL-90R-S that assess exhaustion (item 71) and irritability (item 11).

INSERT FIGURE 1 HERE

Table 2 shows the descriptive statistics of the top 11 important variables.

INSERT TABLE 2 HERE

Partial dependency plots

Figures 2-3 show the marginal effect of the MSES sum score together with either item 71 of the SCL or the duration of fussing/crying documented in the behavioral diaries. In both figures, a plateau effect of MSES can be observed, where values lower than 31 or higher than 34 have little effect. In addition, figure 3 shows a plateau effect for SCL-90R-S Item 71 (everything feels exhausting): Values below the sample mean of 1.7 are indicative of low PS while values above 1.7 are indicative of higher PS.

INSERT FIGURE 2 AND 3 HERE

Figure 3 shows a linear increasing effect of the duration of fussing/crying on PS up until 500 minutes (8.33 hours per day) while the plot slightly dips afterwards and only five participants reported values above 500 minutes.

Partial dependency plots on the relation between the remaining eight important variables and the PSI score are provided in the supplement (Supplemental Digital Content 2).

Discussion

To the best of our knowledge, this study is the first to have explored factors related to mothers' PS in ERD by including multiple measures. We used a ML approach in order to include many differentially scaled and potentially correlated variables in one prediction model.

As expected, mothers in our sample reported much higher average PS compared to a normative sample. Upon analysis of 464 variables involving self-report questionnaires, behavioral diaries, and clinical assessments, we found 11 important predictors for mothers' PS that can be summed up to the following factors: maternal self-efficacy, psychological distress (especially depression and anger-hostility), infant regulatory symptoms, and age-appropriate physical development.

Overall, our results demonstrate that mothers' level of PS in ERD is mainly associated with current problems in the mother-infant dyad, while distal risk and protective factors are less important. Utilizing cross-validation we found that the model would likely generalize well to a similar population. Thus, the identified key variables can be used to select mothers who are at an increased risk for experiencing high PS and to guide treatment of ERD. Below we discuss the important variables and implications of our results in detail.

Maternal self-efficacy

The maternal self-efficacy (MSES) sum score was the most important predictor in the final model and was – as expected – negatively related to PS. The importance of the construct is in line with previous research: Compromised maternal self-efficacy has been described as an important factor in the etiology or perpetuation of ERD,[6] while higher self-efficacy may be ameliorative to PS.[18] Although mothers in our sample on average rated themselves as "good enough" in terms of how effective they experienced themselves across different parenting situations, the range in this scale was broad (table 2) with the observed minimum of 19 points being equivalent to a rating of "not good enough. Mothers with such low expectations are prone to experience high PS.

In addition, we identified incremental effects between low MSES scores and exhaustion (SCL-90R-S item 71, figure 2) and duration of infant fussing and crying (behavioral diary, figure 3). This means for example that if a mother reported low self-efficacy and in addition experienced considerable exhaustion or experienced ≥3 hours of fussing/crying per day, the model predicted significantly more PS compared to mothers who did not fit these criteria.

The MSES item "good at keeping baby occupied" had an additional, albeit less important role in the prediction. On average mothers reported comparably lower self-efficacy regarding this specific parenting situation in contrast to the mean of MSES (table 2). Thus, our results highlight a specific aspect of maternal self-efficacy related to ERD – the self-efficacy mothers experience when successfully occupying their infant. This item might be especially relevant because occupying the child is a parenting task that continually arises throughout the day. Low expectations with this regard seem to predict mothers' daily distress levels.

These results have several implications. Firstly, clinicians should assess and be aware of subtle deviations in the MSES in order to align treatment strategies. Secondly, interventions that promote self-efficacy, especially related to parenting situations involving fussing/crying and occupying the child, and with regard to coping with exhaustion, should be considered. Lastly, we identified a subgroup of mothers who reported high self-efficacy who experienced less PS, despite the challenging conditions they faced. Higher self-efficacy may help in coping with prolonged fussing/crying but also in coping with exhaustion. Future research may focus on this subgroup to investigate conditions under which maternal self-efficacy can be a protective factor for PS.

Mothers' psychological distress

The second set of predictors were maternal psychological distress symptoms experienced during the last week, as was reflected in the SCL-90R-S sum score, the subscales depression and anger-hostility, and two items from these subscales (exhaustion and irritability). Surprisingly, these two items were among the three most important predictors in the dataset. The partial dependency plots further specified nearly linear relations between mothers' exhaustion and irritability with the PSI score (see Supplemental Digital Content 2).

We noticed that mothers in our study compared to a normative sample were more psychologically distressed on average and displayed a high range on the SCL-90R-S sum score. Remarkably, *T*-values of the subscales depression ($T = 60$) and anger/hostility ($T = 62$) indicated

a noticeable higher distress in these domains,[24] suggesting that these are specific vulnerability factors in our sample.

Our results add to the notion that parents who are more depressed experience parenting in ERD as more difficult[38] and moreover specify which emotional aspect of depression is especially relevant to maternal PS in ERD. Accordingly, more exhausted mothers experience parenting as even more stressful compared to less exhausted mothers. It is also likely that depressive symptoms and anger-hostility inhibits parenting skills and thus increases PS, given the studies showing that symptoms are linked to parenting impairments[39]. Meanwhile, it is also plausible that a mother, who experiences more difficulties in parenting, reactively develops symptoms of depression and irritability as a result of helplessness and a lack of self-efficacy. Drawing from our results, clinical assessments and treatment conceptualization for ERD should especially consider these specific symptoms.

While we found that current psychological distress symptoms were an important predictor, maternal lifetime mental illness was not among the critical variables. This result aligns with literature showing that PS was unrelated to prenatal anxiety or depression in no-risk infant samples[15] and may indeed play a subordinate role in parental burden related to ERD.[40] However, several aspects need to be considered: mothers with severe psychological distress were excluded from study participation. Additionally, since we utilized only self-report measures, lifetime mental illness may have been underreported.[41] Both of these factors may have contributed to the low prevalence rate of mental illness, thereby reducing the likelihood that this variable is shown as predictive. Future studies should assess a more representative parent sample utilizing interview-based measures.

Infants' regulatory symptoms

Three variables indicative of infants' regulatory symptoms were important in predicting PS: the duration of fussing/crying as documented by mothers in behavioral diaries, the amount

of clinically assessed regulatory symptoms (sum of symptoms in the interview), and the QCFS sum score. As expected, all variables were negatively related to PS.

Behavioral observations of prolonged fussing and crying came up as the fourth most important variable in our dataset. The importance of this variable, as opposed to other ERD-symptoms, was unexpected, as only 8.9% of infants were diagnosed with persistent excessive crying disorder, while almost all infants were diagnosed with a sleeping disorder. The descriptive statistics indicate an overall high level of combined fussing and crying times with a mean of over three hours and a maximum of 15.39 hours per day (table 2). Although values greater than 8.33 hours per day were infrequent, this result is in itself an important contribution to the literature and warrants further investigation. One possible explanation for the high prevalence in our sample is that different ERD are likely related to fussing and crying. That is, difficult sleep-wake regulation has been associated with difficult temperament and low sensory thresholds, which were in turn related to increased fussing and crying.[42, 43]

These results corroborate previous literature on the adverse effect of prolonged crying on parents' level of perceived burden and physiological reactions in no-risk and risk samples.[44, 40, 45] While it is also likely that higher PS, which renders parents are less effective in soothing their child, contributes to more regulation problems, the literature points to negative effects of dysregulation on parents.[4, 5] Both factors – PS and infants' dysregulation – may exist in a reciprocal relationship with each other, thereby contributing to the perpetuation of ERD.[6] Drawing from our results, especially fussing and crying related to ERD may contribute to this build-up.

In our sample, there was a high comorbidity of ERD with almost 50% of the sample fulfilling diagnostic criteria of more than one diagnosis. Accordingly, the scale "sum of symptoms" covers a large range of up to 35 clinically assessed symptoms of different ERD (table 2). Our results imply that for a mother in our sample, the more symptoms the greater levels of PS, irrespective of the nature and quality of the symptoms or the behavioral area

affected. Independently, the extent of mother-reported infant crying, sleeping and feeding symptoms, co-regulation difficulties, and the related burden, were predictive for the level of PS.

These results highlight the need to utilize multiple measures in order to estimate the association between regulatory symptoms and maternal PS. Behavioral diaries seem to capture important aspects of everyday life that are relevant to PS. Self-report measures may add an important subjective factor to the clinically assessed symptoms. For treatment planning, our results suggest targeting mothers' experience of prolonged and inconsolable fussing/crying in sleeping disorders and comorbid ERD.

Infants' age-appropriate physical development

Mothers' rating of an age-adequate physical development of the child was the least import predictor in the final prediction model. While most of the mothers felt that their child was well developed physically (94.08%, table 2), it seems that having the impression of a "normal" development or not, makes a difference to the extent of PS. While interpreting this result, it is important to discuss that infant age-appropriate developmental performance assessed with the ASQ-3 (e.g., gross-motor development), was unrelated to PS. One explanation of this result is that not actual developmental problems, but the mothers' perception thereof is what makes parenting more or less stressful. Asking mothers about their perception of infant development may be a more valuable question in order to estimate their level of PS.

Limitations

Our results' generalizability is restricted by the relatively homogeneous sample in terms of psychosocial and sociodemographic characteristics. This homogeneity led to close-to-zero variance, leading some variables to be excluded by the algorithm, e.g. unemployment of one parent or both. Additionally, the exclusion criteria of this study likely limited the variance in relevant variables like organic and medical infant risk factors and maternal mental illness. Thus, while we cross-validated all of our models, it is likely that the final model does not generalize

to unselected samples of mothers with infants with ERD. For this reason, results of this study should be interpreted with caution, and will need future replication with more diverse samples and fathers.

Further limitations apply to the instruments used. PRFQ, MSES, and ASQ are not validated in German. The clinical interview utilized is not validated. However, infants' clinical characteristics in our study resemble other clinic and at-risk samples,[13, 3] which speaks to the data's generalizability in this regard.

We used items in our dataset on a single item level in order to maximize specificity and to make suggestions for future item selection. This strategy was further necessitated by the low reliability of some subscales (e.g., PRFQ-IC, SCL-psychoticism). Thus, readers should take care when interpreting our results not to infer an underlying construct from a single item.

Finally, while we assessed several risk factors, the use of cross-sectional assessed data in our study excludes causal data interpretation.

Future research

We showed that ML applied to a dataset stemming from multiple measures, can be utilized to predict a mother's PS. Based on this study, future longitudinal studies may utilize ML for the coverage of additional risk and protective factors (e.g., mental illness of both parents, social support) for PSI levels in both parents. Such investigations allow us to explore causal pathways that consider multiple infant and parent variables and their interactions within a family, and a developmentally sensitive perspective on the factors that contribute to PS in ERD. Future studies with naturalistic samples will lead to even greater generalizability of the findings.

References

1. Skovgaard AM, Houmann T, Christiansen E, et al. The prevalence of mental health problems in children 1 1/2 years of age - The Copenhagen Child Cohort 2000. *J Child Psychol & Psychiat.* 2007;48(1):62–70.
2. Yalçin SS, Orün E, Mutlu B, et al. Why are they having infant colic? A nested case-control study. *Paediatr Perinat Epidemiol.* 2010;24(6):584–596.
3. Postert C, Averbeck-Holocher M, Achtergarde S, Müller JM, Furniss T. Regulatory disorders in early childhood: Correlates in child behavior, parent-child relationship, and parental mental health. *Infant Ment Health J.* 2012;33(2):173–186.
4. Pauli-Pott U, Becker K, Mertesacker T, Beckmann D. Infants with "Colic"—mothers' perspectives on the crying problem. *Journal of Psychosomatic Research.* 2000;48(2):125–132.
5. Lester BM, Zachariah Boukydis CF, Garcia-Coll CT, Hole W, Peucker M. Infantile colic: Acoustic cry characteristics, maternal perception of cry, and temperament. *Infant Behavior and Development.* 1992;15(1):15–26.
6. Papoušek M, von Hofacker N. Persistent crying in early infancy: A non-trivial condition of risk for the developing mother-infant relationship. *Child Care Health Dev.* 1998;24(5):395–424.
7. Winsper C, Bilgin A, Wolke D. Associations between infant and toddler regulatory problems, childhood co-developing internalising and externalising trajectories, and adolescent depression, psychotic and borderline personality disorder symptoms. *J Child Psychol Psychiatry.* 2019.
8. Smarius LJCA, Strieder TGA, Loomans EM, et al. Excessive infant crying doubles the risk of mood and behavioral problems at age 5: Evidence for mediation by maternal characteristics. *Eur Child Adolesc Psychiatry.* 2017;26(3):293–302.

9. Mathiesen KS, Sanson A. Dimensions of early childhood behavior problems: Stability and predictors of change from 18 to 30 months. *Journal of Abnormal Child Psychology.* 2000;28(1):15–31.

10. Touchette E, Petit D, Tremblay RE, Montplaisir JY. Risk factors and consequences of early childhood dyssomnias: New perspectives. *Sleep Med Rev.* 2009;13(5):355–361.

11. Martini J, Petzoldt J, Knappe S, Garthus-Niegel S, Asselmann E, Wittchen H-U. Infant, maternal, and familial predictors and correlates of regulatory problems in early infancy: The differential role of infant temperament and maternal anxiety and depression. *Early Hum Dev.* 2017;115:23–31.

12. van der Wal MF, van Eijsden M, Bonsel GJ. Stress and emotional problems during pregnancy and excessive infant crying. *J Dev Behav Pediatr.* 2007;28(6):431–437.

13. Schmid G, Schreier A, Meyer R, Wolke D. Predictors of crying, feeding and sleeping problems: A prospective study. *Child Care Health Dev.* 2011;37(4):493–502.

14. Fredriksen E, von Soest T, Smith L, Moe V. Parenting stress plays a mediating role in the prediction of early child development from both parents' perinatal depressive symptoms. *Journal of Abnormal Child Psychology.* 2019;47(1):149–164.

15. Leigh B, Milgrom J. Risk factors for antenatal depression, postnatal depression and parenting stress. *BMC Psychiatry.* 2008;8:24.

16. Zwart P, Vellema-Goud MGA, Brand PLP. Characteristics of infants admitted to hospital for persistent colic, and comparison with healthy infants. *Acta Paediatr.* 2007;96(3):401–405.

17. Neece C, Baker B. Predicting maternal parenting stress in middle childhood: the roles of child intellectual status, behaviour problems and social skills. *J Intellect Disabil Res.* 2008;52(12):1114–1128.

18. Bolten MI, Fink NS, Stadler C. Maternal self-efficacy reduces the impact of prenatal stress on infant's crying behavior. *J Pediatr.* 2012;161(1):104–109.

19. Georg A, Schröder P, Cierpka M, Taubner S. Elterliche Mentalisierungsfähigkeit und der Zusammenhang mit elterlicher Belastung bei frühkindlichen Regulationsstörungen. *Praxis der Kinderpsychologie und Kinderpsychiatrie.* 2018;67(5):421–441.

20. Monuteaux MC, Stamoulis C. Machine learning: A primer for child psychiatrists. *J Am Acad Child Adolesc Psychiatry.* 2016;55(10):835–836. doi:10.1016/j.jaac.2016.07.766.

21. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Data mining, inference, and prediction.* New York: Springer; 2009.

22. Zero to Three. *Diagnostic classification: 0 – 3. Diagnostic classification of mental health and developmental disorders of infancy and early childhood. Revised edition (DC: 0 – 3R).* Washington, DC: Zero to Three Press; 2005.

23. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF) (Association of the Scientific Medical Societies). Leitlinien der Deutschen Gesellschaft für Kinder- und Jugendpsychiatrie, Psychosomatik und Psychotherapie für Frühkindliche Regulationsstörungen (Nr. 028/028) [Guidelines of the German Society for Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy for early regulatory disorders]. http://www.awmf.org/leitlinien/aktuelle-leitlinien/ll-liste/deutsche-gesellschaft-fuer-kinder-und-jugendpsychiatrie-psychosomatik-und-psychotherapie.html.

24. Franke GL. *SCL-90®-S. Symptom-Checklist-90®-Standard–Manual.* Göttingen: Hogrefe; 2014.

25. Tröster H. *Eltern-Belastungs-Inventar: EBI; deutsche Version des Parenting Stress Index (PSI) von RR Abidin [Eltern-Belastungs-Inventar: EBI; German version of the Parenting Stress Index (PSI) of R. R. Abidin].* Göttingen: Hogrefe; 2011.

26. Teti DM, Gelfand DM. Behavioral competence among mothers of infants in the first year. The mediational role of maternal self-efficacy. *Child Development.* 1991;62(5):918–929.

27. Ramsauer B, Lotzin A, Mühlhan C, et al. A randomized controlled trial comparing Circle of Security Intervention and treatment as usual as interventions to increase attachment security in infants of mentally ill mothers. Study Protocol. *BMC Psychiatry.* 2014;14:24.

28. Cierpka M. Familienstützende Prävention. In: Cierpka M, ed. *Frühe Kindheit 0-3 Jahre [Early Childhood 0-3 years].* Berlin, Heidelberg: Springer; 2014:523–531.

29. Laucht M, Esser G, Schmidt MH, et al. "Risikokinder": Zur Bedeutung biologischer und psychosozialer Risiken für die kindliche Entwicklung in den beiden ersten Lebensjahren ["Children at risk": The significance of biological and psychosocial risks for child development in the first two years of life]. *Praxis der Kinderpsychologie und Kinderpsychiatrie.* 1992;42(8):275–285.

30. Groß S, Reck C, Thiel-Bonney C, Cierpka M. Empirische Grundlagen des Fragebogens zum Schreien, Füttern und Schlafen (SFS). *Praxis der Kinderpsychologie und Kinderpsychiatrie.* 2013;62(5):327–347.

31. Papoušek M, Rothenburg S, Cierpka M, von Hofacker N. *Regulationsstörungen der frühen Kindheit. CD-basierte Fortbildung [Regulatory disorders in early infancy. CD-based training].* München: Stiftung Kindergesundheit; 2006.

32. Squires J, Twombly E, Bricker D, Potter L. *Ages & Stages Questionnaires.* Baltimore, MD: Brookes Publishing; 2009.

33. van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Soft.* 2011;45(3).

34. R Development Core Team. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing; 2008.

35. Friedman JH. Greedy function approximation: A gradient boosting machine. *Annals of Statistics.* 2001;29(5):1189–1232.

36. Lampa E, Lind L, Lind PM, Bornefalk-Hermansson A. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. *Environ Health.* 2014;13:57.

37. Kessler RC, Aguilar-Gaxiola S, Alonso J, et al. The global burden of mental disorders: An update from the WHO World Mental Health (WMH) Surveys. *Epidemiol Psichiatr Soc.* 2009;18(1):23–33.

38. Pauli-Pott U, Mertesacker B, Beckmann D. Predicting the development of infant emotionality from maternal characteristics. *Develop. Psychopathol.* 2004;16(01):299.

39. Du Rocher Schudlich TD, Norman Wells J, Erwin SEA, Rishor A. Infants' emotional security: The confluence of parental depression, interparental conflict, and parenting. *Journal of Applied Developmental Psychology.* 2019;63:42–53.

40. Radesky JS, Zuckerman B, Silverstein M, et al. Inconsolable infant crying and maternal postpartum depressive symptoms. *Pediatrics.* 2013;131(6):e1857-64.

41. Takayanagi Y, Spira AP, Roth KB, Gallo JJ, Eaton WW, Mojtabai R. Accuracy of reports of lifetime mental and physical disorders: results from the Baltimore Epidemiological Catchment Area study. *JAMA Psychiatry.* 2014;71(3):273–280.

42. Novosad C, Freudigman K, Thoman EB. Sleep patterns in newborns and temperament at eight months: a preliminary study. *J Dev Behav Pediatr.* 1999;20(2):99–105.

43. Touchette E, Côté SM, Petit D, et al. Short nighttime sleep-duration and hyperactivity trajectories in early childhood. *Pediatrics.* 2009;124(5):e985-93.

44. Out D, Pieper S, Bakermans-Kranenburg MJ, van Ijzendoorn MH. Physiological reactivity to infant crying: a behavioral genetic study. *Genes Brain Behav.* 2010;9(8):868–876.

45. Vik T, Grote V, Escribano J, et al. Infantile colic, prolonged crying and maternal postnatal depression. *Acta Paediatr.* 2009;98(8):1344–1348.

TABLE 1

Sample characteristics of infants and their mothers (N = 135)

Variable		***M* / %**	***SD***
Infant age (in months)		8.55	3.10
Mother age (in years)		33.27	4.47
Girls		45.2%	-
First born child		65.2%	-
Mother has high school or higher education		74.8%	-
Mother married		79.3%	-
Mother of German origin		79.3%	
Mother with mental disorder lifetime		14.8%	-
Diagnoses			
	Persistent excessive crying	8.9%	-
	Regulation disorder of sensory processing	44.4%	-
	Feeding disorder	13.3%	-
	Sleeping disorder	95.6%	-
	> 1 diagnoses	48.0%	-
PIR-GAS		74.96	9.76
SCL (GSI)		49.00	34.18
PSI		131.50	31.60

Note. PIR-GAS = Parent-infant relationship general assessment; SCL (GSI) = Global Severity Index of the Symptom-Severity-Check-List-90R-S; PSI = Sum score of the Parenting Stress Index.

TABLE 2

Descriptive statistics of the PSI outcome and the top 11 most important variables for the prediction of PSI

Variable	*M* (%)	*SD*	*Mdn*	min	max	range
Outcome (PSI sum score)	131.50	31.60	131	59	218	159
BD: duration of fussing/crying (M of minutes on 4 days)	189.02	140.88	168.75	11.25	937.5	926.25
CI: sum of symptoms	12.20	5.98	12	3	35	32
MSES sum score	31.7	3.62	31	21	40	19
MSES Item 7: how good at keeping baby occupied?	2.55	0.84	3	1	4	3
PQ: age-appropriate physical development	0.97	0.24	1			2
0 (no)	6 (4.44)					
1 (yes)	127 (94.07)					
2 (uncertain)	2 (1.48)					
QCFS global score	2.21	0.22	2.22	1.63	2.83	1.21
SCL (sum score)	49.00	34.18	41	0	159	159
SCL aggression subscale	4.45	4.42	3	0	20	20
SCL depression subscale	10.99	8.11	9	0	43	43
SCL Item 11: easily irritable	2.05	1.22	2	0	4	4
SCL Item 71: everything feels exhausting	1.70	1.30	2	0	4	4

Note. CI = Clinical interview; MSES = Maternal Self-Efficacy Scale; QCFS = Questionnaire for Crying, Feeding, and Sleeping; SCL = Symptom-Severity-Check-List-90R-S; PQ = Parent-Questionnaire; PSI = Parenting Stress Index; BD = 96-hour behaviour diary.

FIGURE 1

Title: Relative importance of variables from the best predicting model GBM with FS for PSI

Legend: Note. CI = Clinical interview; MSES = Maternal Self-Efficacy Scale; QCFS = Questionnaire for Crying, Feeding, and Sleeping; SCL = Symptom-Severity-Check-List-90R-S; PQ = Parent-Questionnaire; PSI = Parenting Stress Index; BD = 96-hour behaviour diary.

FIGURE 2

Title: Marginal effect of MSES sum score together with the SCL Item 71 on predicted PSI value

Legend: Note. MSES = Maternal Self-Efficacy Scale; SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

FIGURE 3

Title: Marginal effect of MSES sum score together with the duration of fussing/crying (BD) on predicted PSI value

Legend: Note. BD = 96-hour behaviour diary; MSES = Maternal Self-Efficacy Scale; PSI = Parenting Stress Index.

Supplemental Digital Content

Supplemental Digital Content 1

Parent-Questionnaire, 96-hour behavior diary, and clinical interview: scales and items used as predictors

Parent-Questionnaire (110 items)

1) *Sociodemographic information* (33 items)
   infant (5 items): age; gender; nationality; siblings; months apart between siblings
   mother/father (21 items): age; place of birth; nationality; confession; marital status; highest education; professional training; employment status
   family (7 items): living conditions
2) *History of illness* (11 items)
   mother/father (10 items): physical disorder; mental disorder; surgeries; accidents; number of disorders/incidents
   family (relatives) (1 item): disorders/accidents/chronic diseases
3) *Recent significant life events* (5 items)
   divorce/break up; loss of relative; loss of employment; financial difficulties; sudden loss of accommodation/housing
4) *Former miscarriage/abortion* (9 items)
   infertility; duration of involuntary infertility; prior pre-term birth; miscarriage: week of gestation; death of a child
5) *Pregnancy* (11 items)
   planned pregnancy; child desired (mother/father); degree of psychological, social, medical problems/burden during pregnancy; treatment of medical complications in hospital (yes/no; how long); usage of medication during pregnancy; smoking during pregnancy; alcohol consumption during pregnancy
6) *Obstetric history* (6 items)
   week of gestation; hours in labour; type of delivery; other birth complications; subjective burden (mother/father)
7) *Infant medical history* (23 items)

weight at birth; length at birth; head circumference; incubator (yes/no; how long), artificial respiration (yes/no; how long); tube feeding (yes/no; how long); treatment of icterus; onset of problems; asked for help at another service; current treatment; medication; severe or frequent illness of the child: onset, frequency, duration of illness in days; allergies and intolerances; inpatient treatment in hospital: (yes/no), frequency, duration

8) *Infant development* (3 items)

age-appropriate physical development; age-appropriate mental development; age-appropriate social development

9) *Infant social environment* (9 items)

care provided by others: yes/no, since when, type of care, frequency, duration/length, satisfaction; change of caretaker: yes/no, number, age of child

Clinical interview (150 items/scores)

1) *Past regulatory problems* (1 item)

behavioral area affected in the past (categorical)

2) *Persistent excessive crying* (6 items, 1 score)

duration of fussing/crying episodes > 3 hours per day (yes/no);
frequency of fussing/crying episodes > 3 times per way (yes/no);
fussing/crying episodes since at least 3 weeks (yes/no);
lack of success of soothing strategies (yes/no);
episodes more often during the evening (yes/no);
general burden related to persistent excessive crying (0-3);
sum of symptoms

3) *Feeding disorders (DC:0-3R)* (31 items, 6 scores)

Feeding disorder associated with concurrent medical condition (7 items, 1 score)

current medical condition associated with feeding problems (yes/no);
refusal to eat (yes/no);
more distress over course of feeding (yes/no);
fails to gain weight or loses weight (yes/no);
medical management does not fully alleviate the feeding problem (yes/no);
feeding problems since at least 2 months (yes/no);
problems in social responsivity (yes/no);
sum score

Feeding disorder associated with insults to the gastrointestinal tract (5 items, 1 score)

major aversive event or insults (yes/no);

sudden start and fast progression (yes/no);

consistent refusal (yes/no);

trigger of intense distress (yes/no);

food refusal poses an acute or long-term threat (yes/no);

sum score

Sensory food aversions (6 items, 1 score)

consistent refusal of specific foods (yes/no);

onset of food refusal during introduction of a novel type of food (yes/no);

no difficulty with preferred food (yes/no);

refusal to eat and stops eating (yes/no);

specific nutritional deficiencies (yes/no);

problems since at least 1 month (yes/no);

sum score

Infantile anorexia (5 items, 1 score)

lack of interest in food and hunger signals (yes/no);

onset while changing food (yes/no);

significant growth deficiency (yes/no);

refusal to eat adequate amounts of food (yes/no);

problems since at least 1 month (yes/no);

sum score

Feeding disorder of state regulation (4 items, 1 score)

difficulty reaching and maintaining a calm state during feeding (yes/no);

start of difficulties in newborn period (yes/no);

fails to gain weight or loses weight (yes/no);

problems since at least 2 months (yes/no);

sum score

Feeding disorder of caregiver-infant reciprocity (4 items, 1 score)

difficulties in social reciprocity while feeding (yes/no);

primary caregiver ignores feeding or growth problems (yes/no);

significant growth deficiency (yes/no);

exclusion of organic problems or developmental disorder (yes/no);

sum score

4) *Feeding disorder (AWMF)* (10 items, 1 sum score)

more than 45minutes for one feeding episode (yes/no);

less than 2 hours between feeding episodes (yes/no);

growth deficiency (yes/no);

exclusion organic disorder (yes/no);

lack of hunger signals (yes/no);

distraction or forced feeding (yes/no);

age-inappropriate eating behavior (yes/no);

rumination, vomiting (yes/no);

problems to chew, suck or swallow (yes/no);

orofacial sensitivity (yes/no);

sum score

5) *Sleep onset disorder (DC:0-3)* (5 items, 1 score)

time to fall asleep > 30 min. (yes/no);

parent stays in the room until falling asleep (yes/no);

reunions with the parent > 3 times (yes/no);

sleep onset problem episodes 5-7 times during a week (yes/no);

significant difficulties since at least 4 weeks (yes/no);

sum score

6) *Night-waking disorder (DC:0-3)* (5 items, 1 score)

time to fall asleep again > 30 min. (yes/no);

relocation to parental bed (yes/no);

frequency of night-waking during a night > 3 times (yes/no);

night-waking problem episodes 5-7 times during a week (yes/no);

significant difficulties since at least 4 weeks (yes/no);

sum score

7) *Regulation disorder of sensory processing (DC:0-3)* (31 items, 3 scores)

Hypersensitivity (16 items, 1 score)

reacts strongly to sensory stimuli (yes/no);

reacts with aversion to sensory stimuli (yes/no);

avoids strong sensory stimuli (yes/no);

difficulties with postural control and tone (yes/no);

less exploration than expected for age (yes/no);

limited sensory-motor play (yes/no);
general cautious/fearful/avoidant behavioral pattern (yes/no);
restricted range of exploration (yes/no);
fear and clinginess in new situations (yes/no);
distress when routines change (yes/no);
general avoidant behavioral pattern (yes/no);
defiant and avoidant behavior (yes/no);
negativistic behavioral pattern (yes/no);
difficulty adapting to changes in routines/plans (yes/no);
preference for repetition (yes/no);
controlling, compulsive, perfectionistic behavior (yes/no);
sum score

Hyposensitivity (8 items, 1 score)
underreacts to sensory stimuli (yes/no);
lack of responsivity in social interactions (yes/no);
restricted range of exploration (yes/no);
restricted play repertoire (yes/no);
poor motor planning and clumsiness (yes/no);
lack of interest in exploring things or in social interactions (yes/no);
fatigability (yes/no);
withdrawal from stimuli (yes/no);
sum score

Sensory stimulation-seeking/impulsive (7 items, 1 score)
craves for high-intensity sensory stimuli (yes/no);
destructive or high-risk behaviors (yes/no);
high need for motor discharge (yes/no);
impulsive and uncoordinated behavior (yes/no);
seeking constant contact with people and objects (yes/no);
recklessness (yes/no);
general high activity level (yes/no);
sum score

8) *Pervasive regulatory disorder (AWMF)* (10 items, 1 score)
additional behavioral area affected: persistent excessive crying (yes/no);
additional behavioral area affected: night-waking problems (yes/no);

additional behavioral area affected: sleep onset problems (yes/no);
additional behavioral area affected: feeding problems (yes/no);
significant difficulties since at least 4 weeks (yes/no);
significant difficulties on at least 4 days per weeks (yes/no);
symptoms vary in intensity, duration, and frequency (yes/no);
change in behavioral areas affected (yes/no);
symptoms related to specific social interaction partners (yes/no);
dysfunctional interaction patterns (yes/no);
sum score

9) *PIR-GAS score (DC:0-3R)* (1 score)

10) *Biological/organic risk scale* (10 items, 1 score)
10 items of the organic risk scale (Laucht et al., 1992)
sum score

11) *Psychosocial risk scale* (12 items, 1 score)
11 items of the psychosocial risk scale (Laucht et al., 1992)
impact of risk on the child (scale 0-3)
sum score

12) *Emotions and social functioning scale (DC:0-3R)* (4 items, 1 score)
attention and regulation (scale 1-6);
forming relationships/mutual engagement (scale 1-6);
intentional two-way communication (scale 1-6);
complex gestures and problem solving (scale 1-6);
sum score

13) *Sum scores* (7 scores)
sum of symptoms regulation disorders of sensory processing;
sum of symptoms sleep onset and night-waking disorders;
sum of symptoms feeding disorders;
sum of symptoms on axis 1 and persistence excessive crying symptoms;
sum of risk scores;
sum of symptoms on axis 1 and persistence excessive crying symptoms, risk scores, PIR-GAS, and social-emotional functioning
number of diagnosis

96-hour behavior diary (139 items/scores)

1) *Items assessed at each of the 4 days* (29 items*4):

Breast feeding/feeding (minutes, frequency);

Fussing (minutes, frequency);

Crying (minutes, frequency);

Physical contact/carrying (minutes, frequency);

Sleeping during the day (minutes, frequency);

Sleeping at night (minutes);

Sleeping in separate bed (minutes);

Sleeping in parental bed (minutes);

Change of sleeping settings during the night (yes/no);

Time to fall asleep (minutes);

Parental support to fall asleep (yes/no, frequency);

Waking up during the night (frequency);

Parental support to fall asleep after waking up (yes/no, frequency);

Awake during the night (minutes);

Perceived burden related to sleeping behavior (scale 0-3);

Fusses/cries ≥3 hours (yes/no);

Applied soothing strategies (no.);

Success of soothing strategies (scale 0-2);

Duration of applying soothing strategies (minutes);

Perceived burden related fussing/crying (scale 0-3);

Fussing/crying (minutes, frequency)

2) *Scores calculated across 4 days* (23 scores):

Success of soothing strategies (scale 0-2, mean);

Applied soothing strategies (no., sum);

Perceived burden related to sleeping behavior (scale 0-3, mean);

Perceived burden related fussing/crying (scale 0-3, mean);

Parental support to fall asleep (yes/no, mean);

Parental support to fall asleep (yes/no, sum);

Parental support to fall asleep after waking up (yes/no, mean);

Parental support to fall asleep after waking up (yes/no, sum);

Time to fall asleep (minutes, mean);

Sleeping in separate bed (minutes, mean);

Sleeping in parental bed (minutes, mean);

Sleeping during the day (minutes, mean);
Sleeping during the day (frequency, mean);
Breast feeding/feeding (minutes, mean);
Breast feeding/feeding (frequency, mean);
Physical contact/carrying (frequency, mean);
Sleeping at night (minutes, mean);
Awake during the night (minutes, mean);
Waking up during the night (frequency, mean);
Duration of applying soothing strategies (minutes, mean);
Fussing/crying (minutes, mean);
Fussing/crying (frequency, mean);
Fusses/cries ≥3 hours (yes/no, sum)

Supplement B

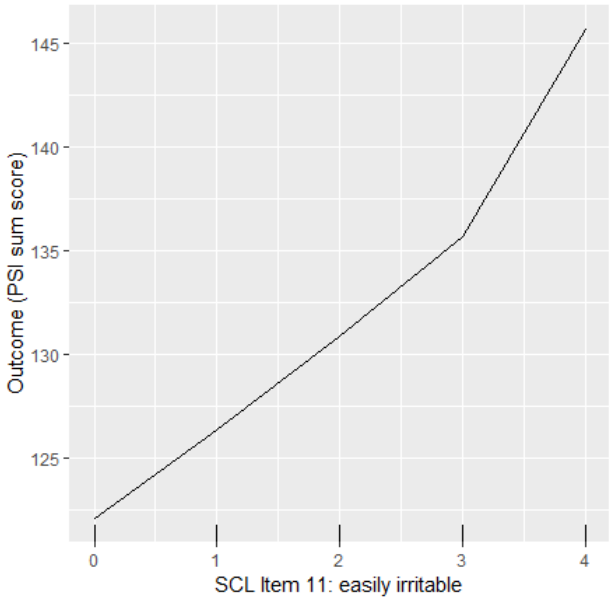


Figure B.1. Partial dependency plot of the SCL Item 11 (easily irritable) on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

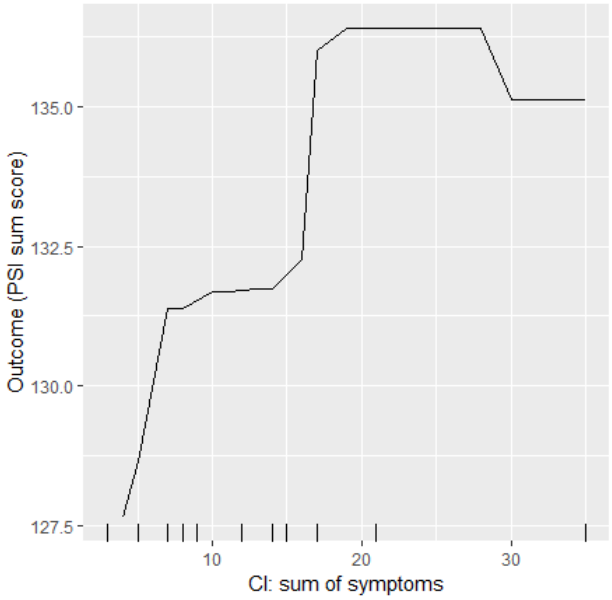


Figure B.2. Partial dependency plot of the CI sum of symptoms score on predicted PSI value. CI = Clinical interview; PSI = Parenting Stress Index.

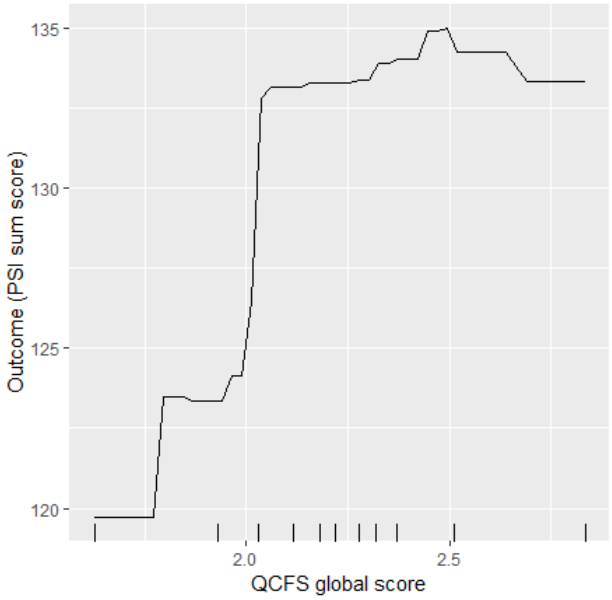


Figure B.3. Partial dependency plot of the QCFS global score on predicted PSI value. QCFS = Questionnaire for Crying, Feeding, and Sleeping; PSI = Parenting Stress Index.

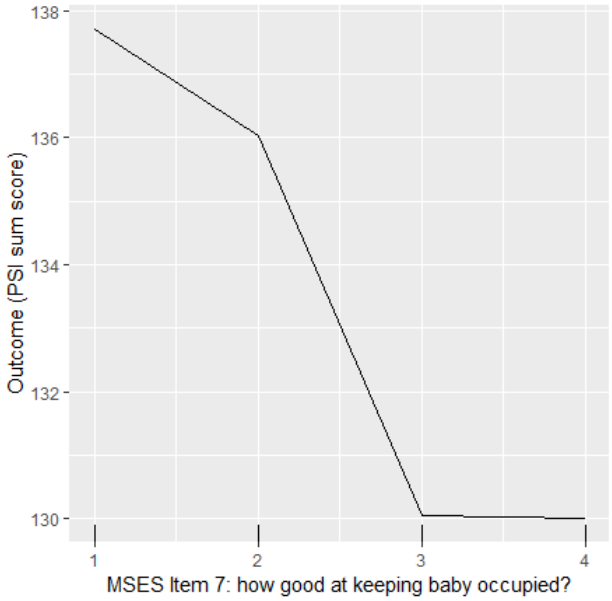


Figure B.4. Partial dependency plot of the MSES Item 7 (how good at keeping baby occupied) on predicted PSI value. MSES = Maternal Self-Efficacy Scale; PSI = Parenting Stress Index.

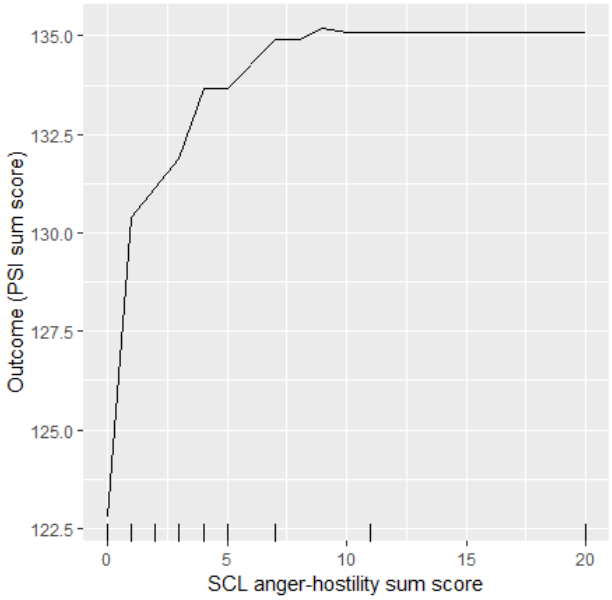


Figure B.5. Partial dependency plot of the SCL subscale anger-hostility on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

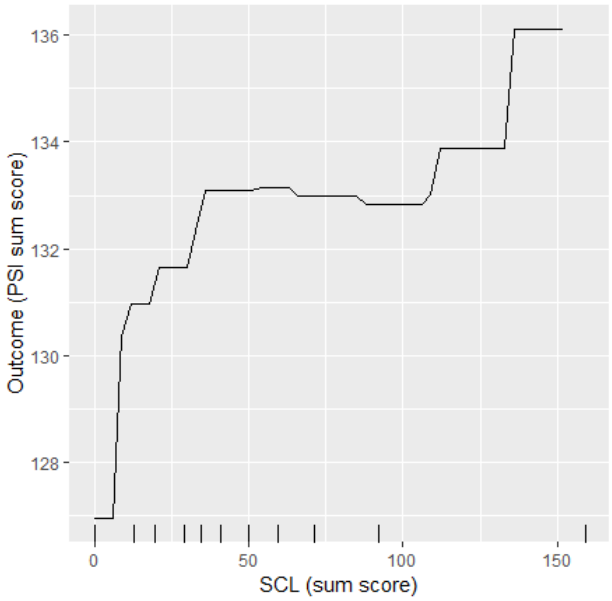


Figure B.6. Partial dependency plot of the SCL sum score on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

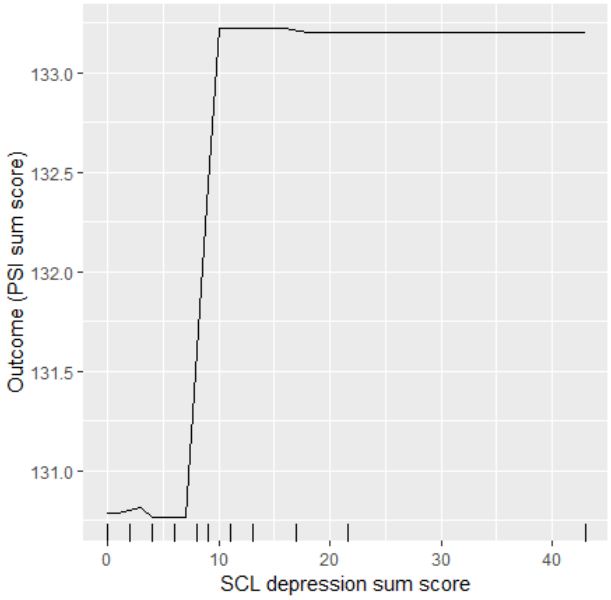


Figure B.7. Partial dependency plot of the SCL subscale depression on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

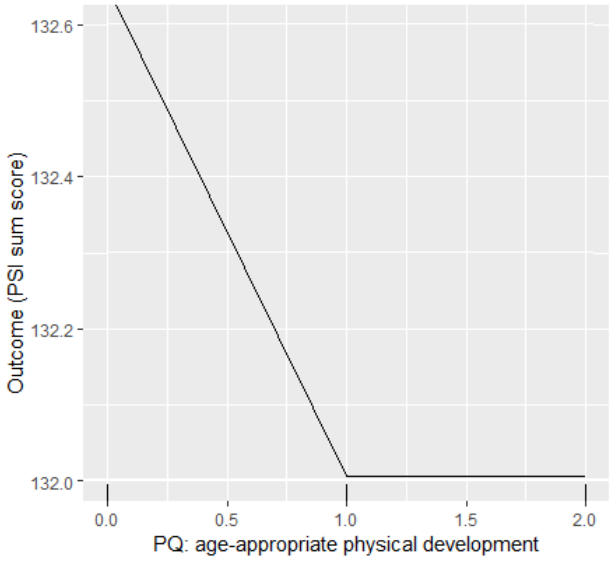


Figure B.8. Partial dependency plot of the PQ item age appropriate physical development on predicted PSI value. PQ = Parent-Questionnaire; PSI = Parenting Stress Index.

Supplemental Digital Content 2

Partial dependency plots on the relation between important variables and the PSI score

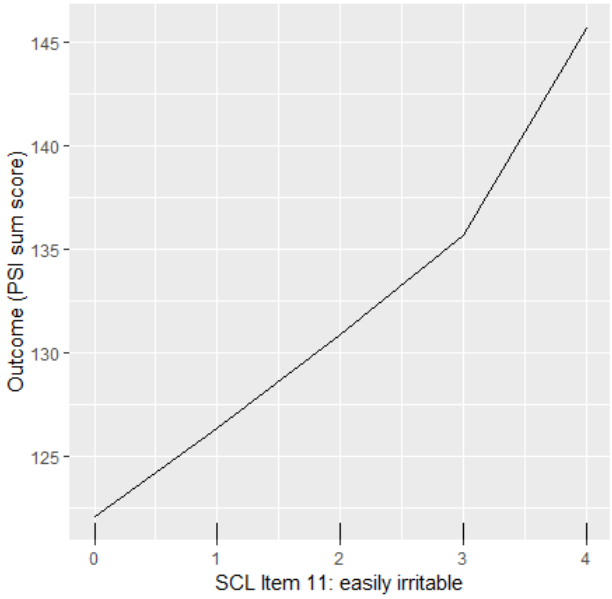


Figure B.1. Partial dependency plot of the SCL Item 11 (easily irritable) on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

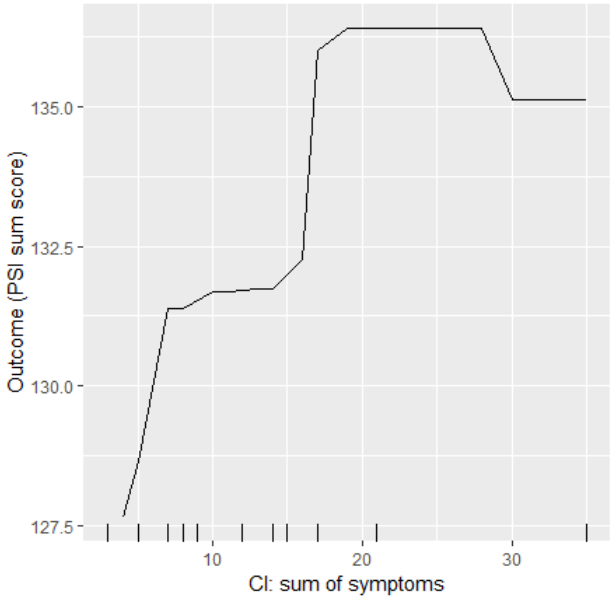


Figure B.2. Partial dependency plot of the CI sum of symptoms score on predicted PSI value. CI = Clinical interview; PSI = Parenting Stress Index.

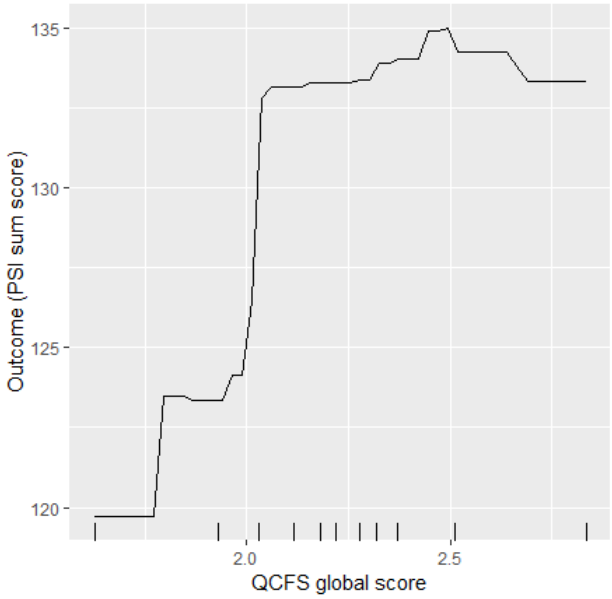


Figure B.3. Partial dependency plot of the QCFS global score on predicted PSI value. QCFS = Questionnaire for Crying, Feeding, and Sleeping; PSI = Parenting Stress Index.

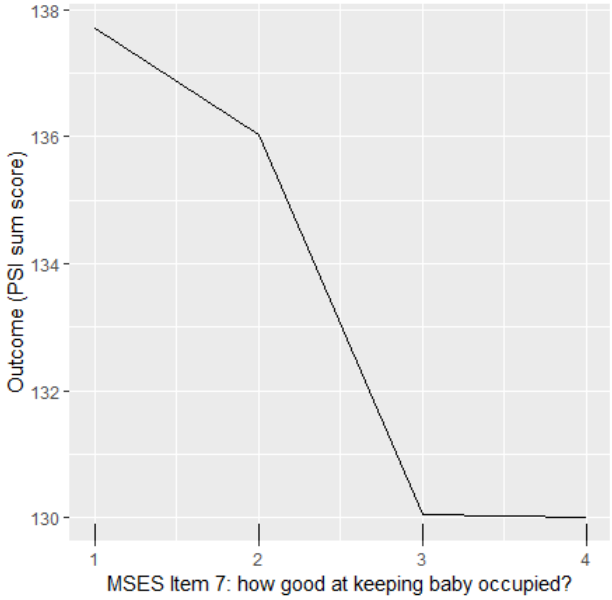


Figure B.4. Partial dependency plot of the MSES Item 7 (how good at keeping baby occupied) on predicted PSI value. MSES = Maternal Self-Efficacy Scale; PSI = Parenting Stress Index.

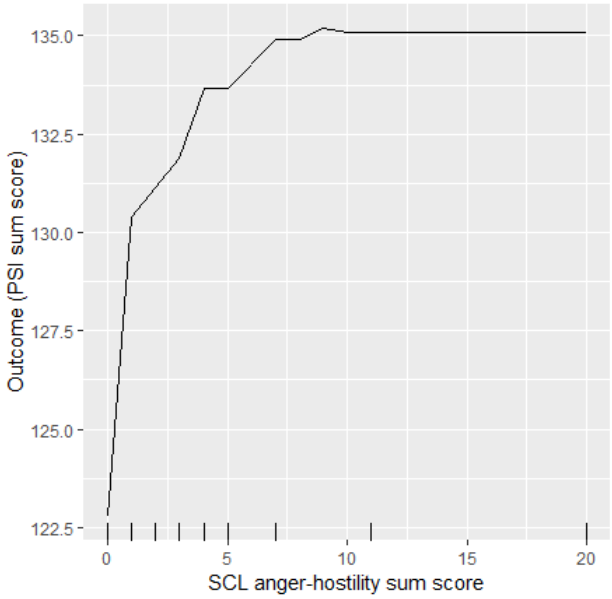


Figure B.5. Partial dependency plot of the SCL subscale anger-hostility on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

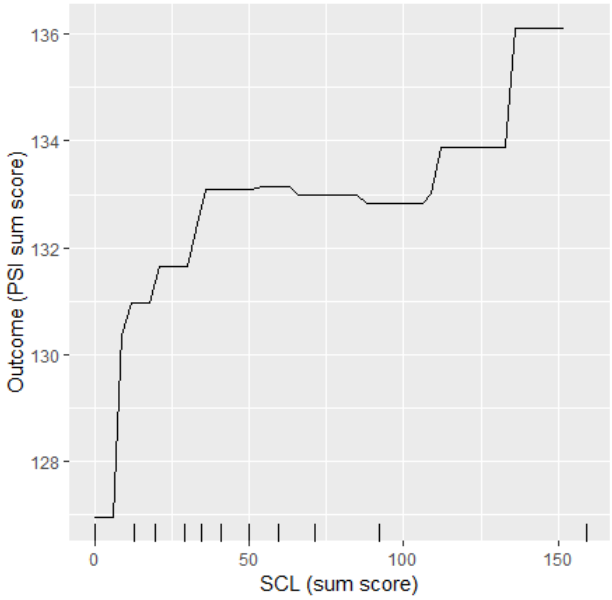


Figure B.6. Partial dependency plot of the SCL sum score on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

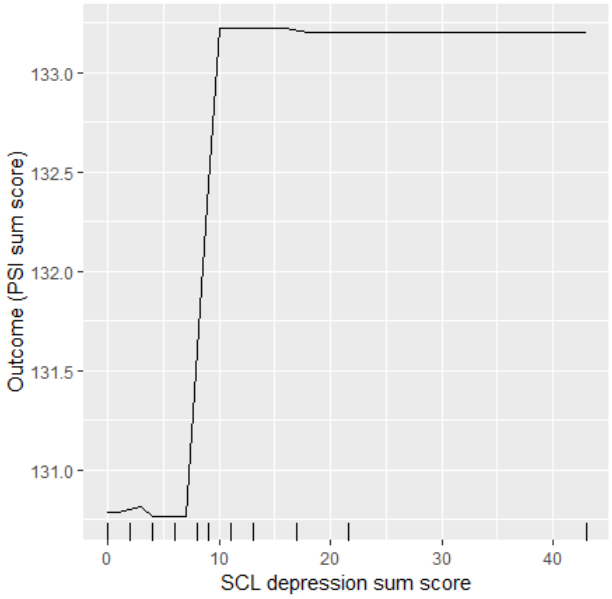


Figure B.7. Partial dependency plot of the SCL subscale depression on predicted PSI value. SCL = Symptom-Severity-Check-List-90R-S; PSI = Parenting Stress Index.

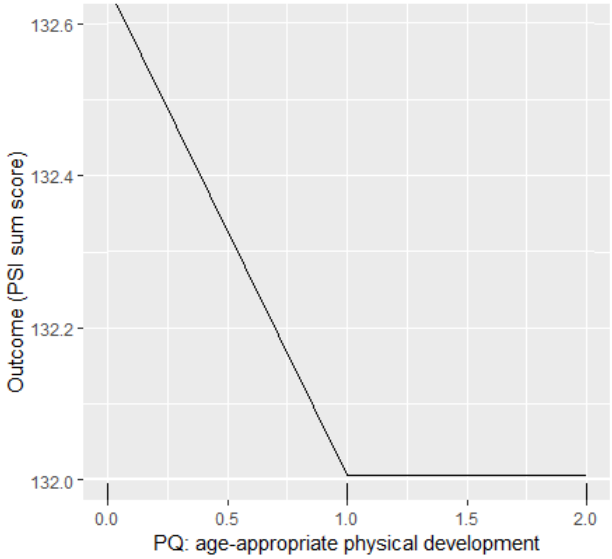


Figure B.8. Partial dependency plot of the PQ item age appropriate physical development on predicted PSI value. PQ = Parent-Questionnaire; PSI = Parenting Stress Index.

Study 3/1

III. **Schröder-Pfeifer, P.**, Talia, A., Volkert, J., & Taubner, S. (2018) Developing an assessment of Epistemic Trust: a research protocol. *Research in Psychotherapy: Psychopathology, Process and Outcome.* IF: -

Declaration of author contributions: Schröder-Pfeifer, P.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Talia, A.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Volkert, J.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

Research in Psychotherapy: Psychopathology, Process and Outcome 2018; volume 21:123-131

Developing an assessment of epistemic trust: a research protocol

Paul Schröder-Pfeifer, Alessandro Talia, Jana Volkert, Svenja Taubner

Institute of Psychosocial Prevention, Heidelberg University Hospital, Heidelberg, Germany

ABSTRACT

Epistemic trust (ET) describes the willingness to accept new information from another person as trustworthy, generalizable, and relevant. It has been recently proposed that a pervasive failure to establish epistemic trust may underpin personality disorders. Although the introduction of the concept of ET has been inspiring to clinicians and is already impacting the field, the idea that there may be individual differences in ET has yet to be operationalized and tested empirically. This report illustrates the development of an Epistemic trust assessment and describes the protocol for its validation. The sample will include 60 university students. The Trier Social Stress Test for Groups will be administered to induce a state of uncertainty and stress, thereby increasing the relevance of information for the participants. The experiment will entail asking information from the participants about their performance and internal states during a simulated employment interview, and then tracking how participants are able to revise their own judgments about themselves in light of the feedback coming from an expert committee. To control for social desirability and personality disorder traits, the short scale for social desirability (Kurzskala Soziale Erwünschtheit-Gamma) and the Inventory of Personality Organization are utilized. After the procedure, the participants will complete an app-based Epistemic trust questionnaire (ETQ) app. Confirmatory Factor Analysis will be utilized to investigate the structure and dimensionality of the ETQ, and ANOVAs will be used to investigate mean differences within and between persons for ET scores by item category. This study operationalizes a newly developed ET paradigm and provides a framework for the investigation of the theoretical assumptions about the connection of ET and personality functioning.

Key words: Epistemic trust; Experiment; Operationalization.

Correspondence: Paul Schröder-Pfeifer, Institute of Psychosocial Prevention, Heidelberg University Hospital, Bergheimer Str. 54, 69115 Heidelberg, Germany.
Tel.: +49.6221.56.38504 - Fax: +49.6221.56.4702.
E-mail: Paul.Schroeder-Pfeifer@med.uni-heidelberg.de

Citation: Schröder-Pfeifer, P., Talia, A., Volkert, J., & Taubner, S. Developing an assessment of epistemic trust: a research protocol. *Research in Psychotherapy: Psychopathology, Process and Outcome, 21*(3), 123-131. doi: 10.4081/ripppo.2018.330

Contributions: all the authors participated in designing the experiment. PS-F did the literature review and first draft of the paper; AT and PS-F provided the introduction; JV and ST revised and edited the manuscript.

Conflict of interest: the authors declare no potential conflict of interest.

Funding: the German psychoanalytic society (Deutsche Psychoanalytische Gesellschaft) provided funding to reimburse the study participants.

Conference presentation: the first draft of the experimental design was presented at the Society for Psychotherapy Research Meeting, Toronto 2017; International Society for the Study of Personality Disorders Meeting, Heidelberg 2017; Deutsche Psychoanalytische Gesellschaft Yearly Meeting 2018; and Society for Psychotherapy Research Meeting, Amsterdam 2018.

Received for publication: 22 August 2018.
Revision received: 20 November 2018.
Accepted for publication: 26 November 2018.

Research in Psychotherapy:
Psychopathology, Process and Outcome 2018; 21:123-131
doi:10.4081/ripppo.2018.330

Introduction

In the past few years, an important shift has occurred in Fonagy, Luyten, Allison, and Campbell views on psychopathology (Fonagy, Luyten, Allison, & Campbell, 2017). Previously (2004), these authors argued that the capacity to reflect on mental states underlying behavior (*i.e.* the capacity to mentalize) is a developmental achievement that arises out of secure attachments, and that mentalizing and secure attachment constitute a source of resilience against psychopathology (Fonagy, 2004). More recently, however, Fonagy, Luyten, and Allison (2015) have proposed that it is disruptions in early social communication - rather than in early attachments or mentalizing *per se* - that lead to subsequent vulnerabilities for psychopathology. Drawing among others from Csibra and Gergely's Natural Pedagogy (Csibra & Gergely, 2009), and from Sperber and Wilson's Relevance Theory (Sperber et al., 2010; Wilson & Sperber, 2012), Fonagy et al. have built the case that psychopathology, insecure attachment, and impaired mentalizing are all linked because they are associated with difficulties in trusting the relevance and generalizability of intentional communication (Fonagy et al., 2017). They refer to this capacity with the term "epistemic trust", and they view its recovery as lying at the heart of any effective psychotherapy.

While these novel views have started to impact clinical and theoretical work (Bateman & Fonagy, 2016; Holmes & Slade, 2017), there is still very little empirical work to

support them. In fact, while the concept of ET has inspired a growing empirical literature in developmental psychology (*e.g.* Corriveau & Harris, 2009; Hagá & Olson, 2017; Harris & Corriveau, 2011), the study of the concept in adolescents, adults and (in particular) clinical populations is still in its infancy. In particular, there is no valid measure of ET available today for adolescents or adults.

The current study attempts to fill this research gap by devising an assessment of epistemic trust (ET) that attempts to translate the theoretical assumptions of the clinically informed ET literature into a valid experimental paradigm. Most of the work in this field up today is theoretical in nature, and further developments in this area of research are likely to depend on methodological advancements related to the measurement of ET. After a brief review of the theoretical framework and empirical literature for this study, in the following we describe the development of our assessment of ET and a protocol for its validation.

Epistemic trust and epistemic vigilance

Learning involves, by definition, some kind of generalization of the import of new information that is learnt on a specific occasion (*i.e.* at a specific time and in a specific place) to novel instances where the information can be used for a different goal or in a different context. Theories of learning usually argue that such generalization relies on statistical procedures that sample multiple episodes (Csibra & Gergely, 2009). Humans, however, can acquire generic knowledge from a single instance in which they gain new information, *i.e.* through intentional communication with a trusted person. For example, from many repeated observations, one may learn that a particular series of movements leads to having one's shoes laced. Yet if the person (*e.g.*, a parent) who is performing those movements does not merely perform the sequence of actions, but performs it *manifestly* for their addressee (*e.g.*, a child) by clearly indicating that this is a demonstration presented to them specifically, they will learn significantly more from the same action than they would from simply observing how it is performed. In other words, by providing information *ostensively* (*i.e.* by indicating an intent to communicate, Sperber & Wilson, 1995), it may suffice one or two demonstrations from a trusted other (*i.e.* a parent) about *e.g.*, "how one ties shoe laces" to transmit information reliably.

Mammal species have developed mechanisms to protect themselves from deception; similarly, humans depend to a large extent on communication with others, which leaves them open to the risk of being misinformed, sometimes intentionally. To ensure that communication remains advantageous, humans must possess a suite of mechanisms for epistemic vigilance (Sperber et al., 2010). However, the human capacity to acquire from others information that has social and cultural significance may rely on a special kind of trust that may be characteristic of the human species.

Csibra & Gergely (2009) have made the claim that human communication is adapted to allow the transmission of generic knowledge between individuals in at least two distinct ways. First, human infants are sensitive by default to ostensive signals that indicate that they are being addressed. Ostensive cues like eye contact, motherese and marked mirroring prepare the interlocutor for information specifically *relevant* to them, thereby increasing the chance of the information being accepted and generalized to other circumstances, interaction partners and situations (Csibra & Gergely, 2009; Egyed, Kiraly, & Gergely, 2013). Second, humans may be biased to interpret ostensive communication as conveying information that is generalizable – *i.e.* have ET.

Epistemic trust, psychopathology, and psychotherapy

Fonagy et al. have drawn from these views to argue about the importance of ET in psychopathology and psychotherapy. ET within an individual is thought to develop in early attachment relationships with primary caregivers (Csibra & Gergely, 2009; Fonagy et al., 2015). In this perspective, personality disorder is seen as descending from a failure to establish ET in early relationships, and identifiable by persistent problems in communication that reveal a lack of trust in interpersonally transmitted information (Allison & Fonagy, 2016; Fonagy et al., 2015; Fonagy & Allison, 2014).

A healthy ET can be described as the capacity to exert appropriate vigilance in the face of possible deceit while maintaining general trust in interpersonally transmitted information (Sperber et al., 2010). On the other hand, the capacity for ET of an individual can be limited in one of two ways. First, an individual might be epistemically hypervigilant (Sperber et al., 2010) or petrified (Fonagy & Allison, 2014), unable to accept information from the outside world, and rigid in their mental states and in behavior. Second, an individual might be epistemically naïve (Sperber et al., 2010), which might lead to a predisposition to being more easily deceived and naïve behavior.

For example, patients with a borderline personality disorder (BPD) have been found to systematically over-attribute hostile intentions to other people (Nicol, Pope, Sprengelmeyer, Young, & Hall, 2013), over-interpret motives of other people (Sharp et al., 2011; Sharp et al., 2013), and broadly speaking misattributing mental states (*e.g.* Daros, Uliaszek, & Ruocco, 2014; Matzke, Herpertz, Berger, Fleischer, & Domes, 2014). Research suggests that patients with BPD consistently perceive the reason for someone's behavior as threatening or at least malevolent and therefore disregard information provided by their social interaction partners, consistent with their view of the social world being generally malevolent. This phenomenon is not only found in BPD but also in other personality disorders (*e.g.* Bateman & Fonagy, 2016; Beck, Davis, & Freeman, 2016; Schnell & Herpertz, 2018). It translates into a rigidity that hinders the normally ongoing process of updating the

self (beliefs about the world and oneself) based on information from the social environment.

ET has also been discussed as a general mechanism of change in psychotherapy. In psychotherapy, interpersonal processes like empathy, mentalization, and the therapeutic alliance may be considered to function as ostensive cues (Csibra & Gergely, 2009; Fonagy & Allison, 2014). The importance assigned to ET seems compatible with most theories of psychotherapy (*e.g.*, cognitive, psychoanalytic, humanistic) because it tackles a human learning process addressed in any therapeutic intervention: the capacity to learn from experience. The feeling of being understood, of finding oneself accurately represented in the mind of another, rekindles ET and thus might reestablish trust in social learning. This is of central importance for the therapy of individuals with epistemic petrification, which normally experience a sense of isolation from the social world due to communicative pathways with others being essentially severed (Fonagy et al., 2015). Over time, in a benevolent social environment, this may also generalize beyond the therapeutic setting as it enables increasingly accurate interpretation of other's mental states (Fonagy et al., 2015; Fonagy & Allison, 2014).

Previous research

While conceptual work on ET promises to advance our understanding of developmental psychopathology and psychotherapy, there is a need for a valid instrument that assesses ET in adolescents and adults and therewith provides an empirical validation for this clinical theory. In devising our ET instrument we have drawn from previous experimental work carried on young children (Corriveau & Harris, 2009; Egyed et al., 2013). In the following paragraphs we describe these earlier studies and then present how we developed our instrument to study ET in adults. Egyed in his experiment (Egyed et al., 2013) sets out to study the mechanism of ET in toddlers. In Egyed's experiment n=48 toddlers aged 18 months were seated across a table with an experimenter. On the table in between the toddler and the experimenter were placed two objects, one blue object to the right and one orange object to the left. In the first condition, the experimenter first smiled at the blue object and then looked disgusted towards the orange object. The experimenter then left the room and a second person entered and asked the toddler to hand her one of the objects. In 31% of the cases, the toddler handed the object preferred by the experimenter. In contrast, in the second condition, where the experimenter established ostensive contact with the toddler by smiling and eyecontact, the toddler handed the second person the object preferred by the experimenter in 69% of the cases. It can be assumed that the toddler generalized the information regarding the preference beyond the dyadic interaction.

The experiment by Corriveau and Harris (2009) with 147 young children at the age of four to five years works similarly. The children were presented with pictures of fantasy animals and had to choose one of two labels for the animals, one provided by the child's mother, the other by a stranger. The fantasy animals were either completely unfamiliar or hybrid animals that were made up of two different animals in proportions of 50/50 or 75/25. With the unfamiliar animals and the 50/50 ones, the mother and the stranger supplied different, yet fitting labels. For the 75/25 animals, the mother labeled the part of the animal that corresponded to the 25% part while the stranger supplied the label that corresponded to the 75% part. In this experiment, epistemic vigilance would correspond to the children choosing the label supplied from the mother for the unfamiliar and 50/50 conditions, and the label of the stranger for the 75/25 condition.

Both experiments assess ET by measuring how new information is processed by the child. For the information to actually be processed by ET, the information has to be relevant Sperber et al. (2010). Gilbert et al. were able to show that information that has no specific relevance to the subject is automatically accepted as truthful, but is not internalized (Gilbert, Krull, & Malone, 1990; Gilbert, Tafarodi, & Malone, 1993). Non-salient information is not relevant for the self on a conscious or unconscious level, accordingly, there is no risk associated in accepting it, as the information is not considered relevant at any point in the future. At the same time, while keeping the processing cost at a bare minimum, it might be evolutionary optimal to accept non-relevant information as true if it was not merely uttered but asserted, as assuming the information was false would require the individual to question the legitimacy of the assertion.

While it is relatively easy to experimentally establish relevance with young children, it is more difficult to create salient material for adults, who have already formed interests and knowledge. For new information to achieve relevance in the context of existing beliefs, one of three conditions has to be met (Sperber & Wilson, 1995): i) Implications arise taking the new information and contextual beliefs together as premises, which are not derivable from neither the context nor the new information alone. These implications are then accepted as new beliefs. ii) The individual has to adjust their confidence in contextually activated beliefs when taking in the new information. iii) The individual's prior beliefs might contradict the new information. Either the new information has to be rejected or the existing beliefs have to be remodeled accordingly (Sperber & Wilson, 1995).

A further challenge is that the majority of experiments that aim to assess ET with children restrict themselves to presenting to participants declarative information (*e.g.* Corriveau & Harris, 2009; Egyed et al., 2013; Hagá & Olson, 2017). While declarative information has the advantage of establishing the *correct* answer to statements and questions, it may fail to touch on the more socially focused aspects of ET in which *correctness* of inherently subjective information like feedback on a performance has to be established within social interactions.

In sum, the relevance of ET in the field of psychotherapy research has substantially grown in recent years. Yet, to the best of our knowledge, there is no valid measure of ET available for adolescents or adults although some are in development (*e.g.* Luyten, 2017; Nolte, 2017). Accordingly, this study aims to develop an experimental paradigm for the assessment of ET that closely relating to its theoretical basis.

Materials and Methods

Participants

Participants will be students of the University of Heidelberg who have voluntarily signed up to participate in studies via an online study participation platform. Students are notified about the platform by e-mail when they first sign up to university. Inclusion criteria are age above 18, able to provide informed consent, and are fluent in the German language. 2424 registered students at the time of the sighting of the recruitment pool were filtered according to the inclusion criteria and the recruited sample was selected randomly by a computer tool build into the platform from a pool of 1737 eligible students (Figure 1).

Development of the epistemic trust assessment

Building on the ET experiments designed by Egyed et al. (2013) or Corriveau et al. (2009) with young children, we designed the epistemic trust assessment (ETA) to control and observe the content and amount of information passed to an individual and the degree to which the individual internalizes and generalizes that information, this way providing an indirect estimate of ET. Based on the results from Gilbert et al. (Gilbert et al., 1990; Gilbert et al., 1993), and also previous tries at operationalizing ET (Luyten, 2017; Nolte, 2017), the ETA is developed with a focus on the relevance of information passed to the participants. Furthermore, as research from business, organizational and cognitive psychology suggests that individuals experiencing stress are more prone to gathering information from external sources to combat the uncertainty resulting from the stress (Driskell & Salas, 1991; Eysenck, Derakshan, Santos, & Calvo, 2007; Starcke & Brand, 2012), the ETA was devised for use in combination with an artificial stressor, the Trier Social Stress Test for Groups (TSST-G) (Dawans, Kirschbaum, & Heinrichs, 2011), to increase the relevance of the information. According to the theoretical conceptualization of ET outlined above, establishing salience of the information for the participants is of outmost importance, as irrelevant information has no consequences for the individual and activation of ET is not necessary.

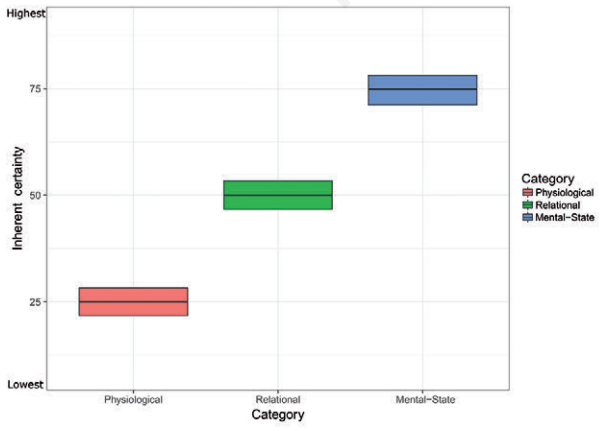


Figure 1. Categories of epistemic trust statements: physiological, relational, mental-state and their inherent certainty (low, medium, high).

In sum, this study aimed to design an experiment utilizing the TSST-G to provide both relevant information to communicate to the participants as well as a context and increased relevance by virtue of providing a stressor. We set out to answer the question, whether or not an experiment can be devised that measures ET and deviations from ET by assessing if participants generalize information supplied to them, given different levels of inherent certainty nested in specific statements. We hypothesize that information can be classified in categories of relative certainty. For the development of the ETA, we differentiated three categories of information that are distinct in terms of their degree of certainty: i) information regarding one's own physiological state (low inherent certainty), ii) regarding relational states (medium inherent certainty), and iii) regarding one's mental state (high inherent certainty). These categories describe three different levels of certainty during the encounter between participant and TSST-G *expert* committee. We assume that specific information about one's own physiological state should be opaque to the individual, and thus have a low inherent certainty. As such, a feedback statement from the expert committee on the individual's heart rate "At the moment your heart rate is around 90 beats per minute." should be difficult to evaluate without the use of technological aides, making questions on physiological states prone to be influenced by feedback. With regard to information on one's mental states is characterized by a high inherent certainty. Assuming that the individual has privileged access to one's mental states, this information should be characterized as high inherent certainty and not be influenced by information from external feedback. Information about relational states can be considered to be of medium inherent certainty as all partners in an interpersonal encounter are considered to have both individual and shared intrapersonal and interpersonal subjective information about the relationship. An individual may have his own judgment on how he is perceived from the outside, but cannot be certain. Consequently, statements regarding relational states should be influenced in a medium way by feedback (Figure 1, Table 1).

Hypotheses

Primary hypothesis

The main hypothesis is that participants adjust their certainty post-feedback according to statement categories and not independent of them. This is assuming a normative sample of participants with healthy epistemic vigilance.

$H0_1$: The participants adjust their certainty post-feedback independent of statement category.
$H1_1$: The participants adjust their certainty post-feedback dependent on category, with most change in the physiological category and least change in the mental states category.

Secondary hypothesis

The secondary hypothesis addresses the relationship between BPD traits and ET. Fonagy et al. (Fonagy et al., 2015; Fonagy & Allison, 2014) conceptualize BPD with the loss of epistemic vigilance tending towards epistemic hypervigilance or equivalent *epistemic petrification*. Accordingly, it is hypothesized that participants with BPD traits adjust their judgments post-feedback significantly less than participants without BPD traits.

$H0_2$: Participants with BPD traits according to the Inventory of Personality Organization (IPO-16) cut-off values adjust their certainty post-feedback the same as participants without BPD traits.
$H1_2$: Participants with BPD traits adjust their certainty post-feedback by significantly less then participants without BPD traits.

Assessment of epistemic trust

Epistemic trust questionnaire

The epistemic trust questionnaire (ETQ) is a self-report questionnaire in app form for the indirect assessment of ET following the ETA. The questionnaire consists of three parts. In the first part, the participants have to rate, according to the 3 certainty categories, their physiological state, their mental states during the TSST-G, and their relational state (*e.g.*, i) "Do you think, your blood pressure (in mmHg) was high or low during the experiment?", ii) "Were you bored during the interview?", iii)"Do you think you came across as motivated?"), and, more importantly, how certain they are in making their judgement. In the second part, the participants are presented with a standardized, computer-generated feedback they think was given to them by the committee, on all of the statements they answered during step one. Finally, in the third step, the participants are asked to re-rate their certainty for the items answered during the first step, taking into account the new information. The items in the first and third step all entail a rating of certainty on a scale of 0 to 100 as well as a binary rating of valence ("Yes/No", "High/Low", *etc*; Figure 2).

The feedback is computer-generated in order to be standardized and is in accordance with the participant's valence rating in exactly half of the questions, as not to introduce a bias on over- or under-agreement. The ET score is operationalized as the difference in certainty from step one to step three, relative to item category. Epistemic vigilance is associated with big changes towards more certainty in the physiological items, medium changes in either direction in the relational items, and no change or small changes in either direction in the mental states items. This operationalization exemplifies epistemic vigilance as a construct of balance that should prompt individuals to internalize and accept information where it is meaningful for them and certainty about their judgment should be low (low certainty item category physiological state). Accordingly individuals should distrust and therefore not internalize information where it is unlikely to meaningfully update their prior knowledge (high certainty item category mental state). Epistemic hypervigiliance is

Table 1. Inherent certainty categories, example items and predisposition to change of the epistemic trust questionnaire.

Category	Example item	Inherent certainty	Predisposition to change
Physiological	"Was your pulse, on average, below or above 97 during the experiment?"	Low	High
Relational	"Do you think you came across as friendly or unfriendly during the experiment?"	Medium	Medium
Mental-State	"Did you feel anxious during the experiment?"	High	Low

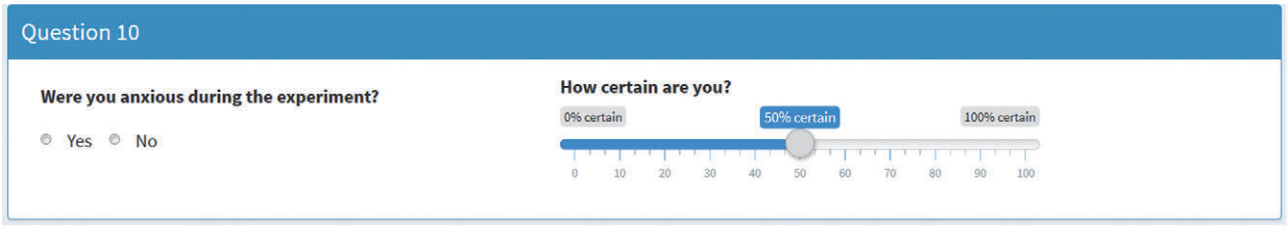


Figure 2. Sample question from the epistemic trust questionnaire.

associated with no or small changes in either direction independent of item category, while epistemic naïveté is associated with big changes towards more certainty independent of item category.

A possible effect known from research on metacognitive phenomenon that might interfere with our hypothesis on how ET is operationalized by the experiment is the so called hypercorrection effect (*e.g.* Butterfield & Metcalfe, 2001; Metcalfe & Finn, 2012). This effect describes a tendency to more easily correct apparently wrong statements that were of high prior certainty as opposed to low prior certainty. This might lead to participants overcorrecting statements with high inherent certainty, such as from the relational and mental states category. However, while this effect has not yet been thoroughly examined for non-declarative information, and research suggests that participants have to be relatively sure that the alternative statement provided to them is correct feedback (Metcalfe & Finn, 2011). In the face of non-declarative information like the feedback provided by the committee in this study, it seems unlikely that this effect applies for any of the categories except for the physiological information, since both relational and mental state information is inherently subjective and can thus never be entirely *correct*.

Social stress test

The Trier Social Stress Test for Groups (TSST-G) (Dawans, Kirschbaum, & Heinrichs, 2011) is a standardized experiment for the reliable induction of moderate social stress (Dawans et al., 2011). The TSST-G is the group version for up to six participants of the original paradigm by Foley and Kirschbaum (2010). The six participants take part in a fabricated job interview combined with an arithmetic task in front of a panel of *experts*. During the interview and the arithmetic tasks, participants cannot see each other, are instructed that they can be called upon at any time in a random order and are being filmed by two cameras. The *expert* panel is instructed to stress the participants by interrupting participants during the interview with questions, if they speak too fluent or too slow as well as prompting them to calculate faster. One *expert* member is the active one, interrupting the participants and asking questions, while the other is appearing to take notes on a laptop for the appearance that data actually utilized. This is a slight modification of the original procedure where the other *expert* member is completely passive. The TSST-G has been shown to reliably induce a robust increase in the activation of the hypothalamic-pituitary-adrenal stress system (Boesch et al., 2014; Kirschbaum, Kudielka, Gaab, Schommer, & Hellhammer, 1999; Kirschbaum, Pirke, & Hellhammer, 1993; Leder, Hausser, & Mojzisch, 2013).

Assessment of social desirability

The Short Scale Social Desirability-Gamma (Kurzskala Soziale Erwünschtheit-Gamma; KSE-G) (Kemper, Beierlein, Bensch, Kovaleva, & Rammstedt, 2012) is an economic measure for the assessment of social desirable behavior (Paulhus, 2015). The scale measures aspects of social desirability associated with a moralistic bias to deny unwanted impulses and to appeal unrealistically positive in the eyes of others. The participants rate six items describing social behavior (*i.e.* "When in an argument, I always stay factual and objective") on a 5-point Likert scale ranging from "does not apply at all" to "applies fully". The authors report satisfactory internal consistency and high factorial and content validity of the instrument (Kemper et al., 2012).

Assessment of personality functioning

The 16-Item-Version IPO-16 (Zimmermann et al., 2013) is a self-report measure to assess personality functioning based on Kernberg's model of borderline personality organisation with regard to identity diffusion, primitive psychological defenses and reality testing. The items are rated on a 5-point Likert scale ranging from "never applies" to "always applies". The authors report good internal constancy (α=.85) and good discriminant, as well as convergent validity (Zimmermann et al., 2013) and also report cut-off values.

In the present study, an app version of both the KSE-G and the IPO-16 was utilized using RShiny (Chang, Cheng, Allaire, Xie, & McPherson, 2017).

Procedure

Participants were sent an email with an outline of the experiment procedure and information regarding the place and date of their experiment session. At arrival on the experiment site, participants were provided detailed information about the type of data assessed in the experiment, the procedure of the assessment, their benefits in participating in the study, as well as contacts for further information and assurance that they could drop out of the experiment at any point in time. However, the underlying aim of the study was obscured in the information material and instead the study's aim was described as exploring the relationship between stress and personality, as well as physiological attributes. After receiving informed consent, the participants were asked to complete both the IPO-16 and KSE-G before undergoing the TSST-G as per protocol (Dawans et al., 2011). The only deviation from the standard protocol was the admission of only four participants at a time, compared to the six from the validation study (Dawans et al., 2011), as the premises did not allow for more participants at one timepoint. After the TSST-G, the ETQ was administered. Finally, the participants were debriefed about the aim of the study and compensated with 10€.

Ethics

The trial received ethical approval from the ethics committee of the Medical Faculty of the University of

Heidelberg, Germany (reference number: S-272/2017). The trial will be conducted in accordance with the European General Data Protection Regulation at all times. Participants will be identified by a study specific participant number during the experiment and anonymized at data aggregation. Names and any other identifying detail will not be included in any study data electronic file. In case sample sizes are very small (subgroups n>20), extra care will be taken by scaling the only personal variable, age, to mean 0 and standard deviation 1, to ensure that individual participants cannot be identified.

Data analysis

A priori estimation of the effect size between the statement categories for this study is not possible, as to our knowledge empirical data on the differences in certainty of retrospective assessments of statements of physiological, relational, and inner states is not available. Therefore, we chose to calculate power based on a medium effect of f^2=.25 between the categories, as a smaller effect could be the result of a flawed conceptualization of the paradigm. An *a priori* power analysis using GPower 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007), using f^2=.25 as effect size, with an alpha of α=.05 and a power of β=.90, resulted in a sample size of n=54. Assuming a drop-out rate of 10% for participants withholding their data for analysis after debriefing, n=60 participants are to be recruited.

In the analysis of the primary hypothesis, the mean certainty ratings post-feedback per category are tested in a two-sided ANCOVA, controlling for gender and a major in psychology, since experience in psychological experiment design might undermine the relevancy aspect of the paradigm for psychology majors. Since all questionnaires utilized in the study are in app form, a forced answer format was chosen to achieve complete data for all participants with no missing values. R (version 3.4.1, R Development Core Team, 2008) is used in all statistical analyses.

Discussion and Conclusions

The described protocol for the validation of a new ET assessment aims to establish a comprehensive and theoretically grounded operationalization of ET in adults. Such new assessment method could provide support for a theory of personality disorder as a failure of communication between the individual and the social environment. It might also prove useful to measure ET pre- and post therapy to study probable predictors of therapeutic outcome. Additionally, being able to reliably measure ET might help disentangle ET, attachment, and mentalizing, three concepts that have historically been hard to separate because they tend to explain similar phenomena on a different level but are also closely related theoretically (*e.g.* Fonagy et al., 2015). Measuring all three constructs in one sample and mapping the relationships between them, ideally with an indicator of severity of personality disorder, ranging from normative to pathological, could provide a valuable empiric underpinning for future research in this field.

Despite these advantages, a number of potential limitations in our assessment need to be addressed. First, given the design of our procedure, its repetition may result in a loss of salience of the information provided and therefore in a lack of relevance. This is particularly unfortunate because repeating the procedure would be needed when attempting to apply it to the study of change, for example in psychotherapy research. In general, our procedure necessarily demands considerable time both from patients and therapist, which limits its applicability. Also, as there are no current alternative measures for ET it is difficult to externally validate the current paradigm except by using theoretically opposing constructs such as a diagnosis of Antisocial Personality Disorder or BPD with which ET should be negatively correlated.

However, if our paradigm will be successfully tested, it will provide the basis for designing more cost- and time-effective measures of ET. For example, a possible adaptation could investigate whether it can be operationalized without the stress inducing component (TSST-G), or whether the presence of a committee (but no job interview or arithmetic task) provides enough salience for the activation of ET. This could prove to be a viable step between an economically viable questionnaire but potentially limited validity and the very time consuming procedure outlined in this study. Another alternative would be to replace the rather rigorous TSST-G with a stressor such as the socially evaluated cold-pressor test (Minkley, Schröder, Wolf, & Kirchner, 2014; Schwabe, Haddad, & Schachinger, 2008). In this procedure, participants are exposed to a physical stressor, as they have to immerse their hand in ice water while they also are continuously observed and evaluated. This procedure could be adapted to include a more pronounced social evaluation aspect that makes it clear to the participants that the *expert* present during the experiment is evaluating them and to use this feedback akin to how the feedback from the committee is used in the present rendition of the ETA. Furthermore, this procedure could be adapted to further investigate the different types and role of ostensive cues in an adult population as well as to investigate the interaction with different psychopathologies.

References

Allison, E., & Fonagy, P. (2016). When is truth relevant? *The Psychoanalytic Quarterly*, *85*(2), 275-303. doi:10.1002/psaq.12074

Bateman, A., & Fonagy, P. (2016). *Mentalization Based Treatment for Personality Disorders* (2nd ed.). Oxford: Oxford University Press.

Beck, A. T., Davis, D. D., & Freeman, A. (2016). *Cognitive therapy of personality disorders* (3rd ed.). New York, London:

The Guilford Press.
Boesch, M., Sefidan, S., Ehlert, U., Annen, H., Wyss, T., Steptoe, A., & La Marca, R. (2014). Mood and autonomic responses to repeated exposure to the Trier Social Stress Test for Groups (TSST-G). *Psychoneuroendocrinology*, *43*, 41-51. doi:10.1016/j.psyneuen.2014.02.003
Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491-1494. doi: 10.1037/0278-7393.27.6.1491
Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). Shiny: Web Application Framework for R. R package version 1.0.3. Available from https://CRAN.R-project.org/package=shiny
Corriveau, K., & Harris, P. L. (2009). Choosing your informant: weighing familiarity and recent accuracy. *Developmental Science*, *12*(3), 426-437. doi: 10.1111/j.1467-7687.2008.00792.x
Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., ... Rosnay, M. de. (2009). Young children's trust in their mother's claims: longitudinal links with attachment security in infancy. *Child Development*, *80*(3), 750-761. doi: 10.1111/j.1467-8624.2009.01295.x
Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148-153. doi: 10.1016/j.tics.2009.01.005
Daros, A. R., Uliaszek, A. A., & Ruocco, A. C. (2014). Perceptual biases in facial emotion recognition in borderline personality disorder. *Personality Disorders*, *5*(1), 79-87. doi: 10.1037/per0000056
Dawans, B. von, Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology*, *36*(4), 514-522. doi: 10.1016/j.psyneuen.2010.08.004
Driskell, J. E., & Salas, E. (1991). Group decision making under stress. *Journal of Applied Psychology*, *76*(3), 473-478. doi: 10.1037/0021-9010.76.3.473
Egyed, K., Kiraly, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*(7), 1348-1353. doi: 10.1177/0956797612471952
Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion (Washington, D.C.)*, *7*(2), 336-353. doi: 10.1037/1528-3542.7.2.336
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175-191. doi: 10.3758/BF03193146
Foley, P., & Kirschbaum, C. (2010). Human hypothalamus-pituitary-adrenal axis responses to acute psychosocial stress in laboratory settings. *Neuroscience and Biobehavioral Reviews*, *35*(1), 91-96. doi: 10.1016/j.neubiorev.2010.01.010
Fonagy, P. (2004). *Affect regulation, mentalization, and the development of the self*. New York NY: Other Press.
Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy (Chicago, Ill.)*, *51*(3), 372-380. doi: 10.1037/a0036505
Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: a new conceptualization of borderline personality disorder and its psychosocial treatment. *Journal of Personality Disorders*, *29*(5), 575-609. doi: 10.1521/pedi.2015.29.5.575
Fonagy, P., Luyten, P., Allison, E., & Campbell, C. (2017). What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personality Disorder and Emotion Dysregulation*, *4*, 9. doi: 10.1186/s40479-017-0062-8
Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*(4), 601-613. doi: 10.1037/0022-3514.59.4.601
Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65*(2), 221-233. doi: 10.1037/0022-3514.65.2.221
Hagá, S., & Olson, K. R. (2017). Knowing-it-all but still learning: Perceptions of one's own knowledge and belief revision. *Developmental Psychology*, *53*(12), 2319-2332. doi: 10.1037/dev0000433
Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366*(1567), 1179-1187. doi: 10.1098/rstb.2010.0321
Holmes, J., & Slade, A. (2017). Attachment in therapeutic practice. Sage.
Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G)*. Available from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-339589
Kirschbaum, C., Kudielka, B. M., Gaab, J., Schommer, N. C., & Hellhammer, D. H. (1999). Impact of gender, menstrual cycle phase, and oral contraceptives on the activity of the hypothalamus-pituitary-adrenal axis. *Psychosomatic Medicine*, *61*(2), 154-162. doi: 10.1097/00006842-199903000-00006
Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' - A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28* (1-2), 76-81.
Leder, J., Hausser, J. A., & Mojzisch, A. (2013). Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology*, *38*(9), 1503-1511. doi: 10.1016/j.psyneuen.2012.12.016
Luyten, P. (2017). *Epistemic trust and BPD: An experimental approach.* Heidelberg, DE: ISSPD.
Matzke, B., Herpertz, S. C., Berger, C., Fleischer, M., & Domes, G. (2014). Facial reactions during emotion recognition in borderline personality disorder: a facial electromyography study. *Psychopathology*, *47*(2), 101-110. doi: 10.1159/000351122
Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: did they know it all along? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(2), 437-448. doi: 10.1037/a0021962
Metcalfe, J., & Finn, B. (2012). Hypercorrection of high confidence errors in children. *Learning and Instruction*, *22*(4), 253-261. doi: 10.1016/j.learninstruc.2011.10.004
Minkley, N., Schröder, T. P., Wolf, O. T., & Kirchner, W. H. (2014). The socially evaluated cold-pressor test (SECPT) for groups: Effects of repeated administration of a combined physiological and psychological stressor. *Psychoneuroendocrinology*, *45*, 119-127. doi: 10.1016/j.psyneuen.2014.03.022
Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social judgement in borderline personality disorder. *PloS One*, *8*(11), e73440. doi: 10.1371/journal.pone.0073440

Nolte, T. (2017). *Epistemic trust in adolescents and BPD patients: Two experimental approximations.* Heidelberg, DE: ISSPD.

Paulhus, D. L. (2015). Socially desirable responding: The evolution of a construct. In H. Braun, Jackson, D.N., & D. E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement* (pp 49-69). London: Routledge.

R Development Core Team. (2008). *A language and environment for statistical computing*. Vienna, Austria.

Schnell, K., & Herpertz, S. C. (2018). Emotion regulation and social cognition as functional targets of mechanism-based psychotherapy in major depression with comorbid personality pathology. *Journal of Personality Disorders*, *32*(Supplement), 12-35. doi: 10.1521/pedi.2018.32.supp.12

Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology*, *33*(6), 890-895. doi: 10.1016/j.psyneuen.2008.03.001

Sharp, C., Ha, C., Carbone, C., Kim, S., Perry, K., Williams, L., & Fonagy, P. (2013). Hypermentalizing in adolescent inpatients: treatment effects and association with borderline traits. *Journal of Personality Disorders*, *27*(1), 3-18. doi: 10.1521/pedi.2013.27.1.3

Sharp, C., Pane, H., Ha, C., Venta, A., Patel, A. B., Sturek, J., & Fonagy, P. (2011). Theory of mind and emotion regulation difficulties in adolescents with borderline traits. *Journal of the American Academy of Child and Adolescent Psychiatry*, *50*(6), 563. doi: 10.1016/j.jaac.2011.01.017

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359-393. doi: 10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (1995). Relevance. Communication and cognition. Oxford, Cambridge, MA: Blackwell Publishers Ltd.

Starcke, K., & Brand, M. (2012). Decision making under stress: a selective review. *Neuroscience and Biobehavioral Reviews*, *36*(4), 1228-1248. doi: 10.1016/j.neubiorev.2012.02.003

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge: Cambridge University Press.

Zimmermann, J., Benecke, C., Hörz, S., Rentrop, M., Peham, D., Bock, A., ... Dammann, G. (2013). Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). *Diagnostica*, *59*(1), 3-16. doi: 10.1026/0012-1924/a000076

Study 3/2

IV. **Schröder-Pfeifer, P.**, Georg, A.K., Talia, A., Volkert, J., Ditzen, B., & Taubner, S. (Under review) The Epistemic Trust Assessment (ETA) – An experimental measure of Epistemic Trust. *Psychoanalytic Psychology.* IF: 0.958

Declaration of author contributions: Schröder-Pfeifer, P.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, Funding Acquisition **Georg, A.K.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Talia, A.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Volkert, J.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Ditzen, B.:** Conceptualization, Writing – Review & Editing Resources **Taubner, S.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Supervision

The Epistemic Trust Assessment – An experimental measure of Epistemic Trust

Paul Schröder-Pfeifer[1,2], Anna K. Georg[1,2], Alessandro Talia[1], Jana Volkert[1], Beate Ditzen[3] and Svenja Taubner[1,2]

[1]Institute for Psychosocial Prevention, Center for Psychosocial Medicine, University Hospital Heidelberg, Heidelberg, Germany

[2]Institute of Psychology, Ruprecht-Karls-University, Heidelberg, Germany

[3]Institute of Medical Psychology, University Hospital Heidelberg, Heidelberg, Germany

Author Note

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Paul Schröder-Pfeifer, Bergheimer Str. 54, 69115 Heidelberg, Germany, Phone: +49 (0)6221 5638504, E-Mail: Paul.Schroeder-Pfeifer@med.uni-heidelberg.de

Abstract

Epistemic Trust (ET) describes an individual's trust in the relevance of interpersonally transmitted information. While this concept increasingly informs theories of communication and psychopathology as well as psychoanalytic change theory, there currently exists no rigorous way of measuring ET. This study describes an experimental paradigm for assessing ET. We designed the epistemic trust assessment (ETA) procedure in which we first utilized the Trier Social Stress Test for Groups (TSST-G), which asks participants to engage in public speaking and mental arithmetic in front of two evaluators and other experimental subjects. Next, the subjects were individually administered a questionnaire, which asked questions about subjects' own behavior and overall performance during the interview. Participants were then given a standardized feedback about their behavior and performance, which included information about aspects in which the evaluators were "trustworthy informants" (e.g., subjects' objectively measured physiology) and "untrustworthy informants" (e.g., subjects' mental states), and they were then asked if they wanted to revise their previous answers. ET was operationalized as the extent to which participants were able to adequately modify their perspective on the basis of evaluators' trustworthy feedback. We controlled for social desirability and personality disorder traits using the Short Scale for Social Desirability (KSE-G) and the short form of the Inventory of Personality Organization (IPO-16). The results confirmed our hypothesis. A majority of participants endorsed trustworthy feedback and rejected untrustworthy feedback. The ETA can be used as an internally validated measure of ET. Future studies validating the ETA in a clinical population are warranted.

Keywords: Epistemic Trust, Experimental Assessment, Communication, Trier Social Stress Test, Psychotherapy Process.

Word count: 7888

During the last century, psychoanalysis and psychoanalytic therapies have been guided by one of two fairly different paradigms of therapeutic action (see e.g. Jones, 2000). According to the first one, therapeutic change occurs when patients gain a better understanding of what unconsciously drives their behavior (Freud, 1923). According to the second one, change occurs because patients experience a new type of human relationship with the therapist, which helps them revise their pathogenic interpersonal expectations (Strachey, 1934; Alexander & French, 1946). Today, the debate about the therapeutic processes that may be most transformative – insight (see e.g., Kernberg, et al. 2008) or new relational experiences (see e.g., BCPSG, 2005) - is ongoing, with several authors who have championed the need to integrate the two approaches (see e.g., Mitchell, 1999; Bateman & Fonagy, 2016).

The recent use of the concept of epistemic trust (ET) in clinical psychology (Fonagy & Campbell, 2015) seems to hold the potential of reconciling the opposing factions of this debate. ET has been defined as the expectation that *interpersonally communicated knowledge* may be true and relevant (Fonagy, et al., 2016). Fonagy et al. (2016) have proposed that ET can be influenced by unconscious, developmentally-based expectations about the competence and benevolence of others. These expectations seem to be especially mistrustful in patients with personality disorders. In this perspective, therapy should aim at transforming patients' unconscious expectations that social information is misleading or irrelevant and re-open a possibility for learning from others. Namely, Fonagy et al. view therapy as pursuing three related tasks: (1) helping patients acquire relevant *knowledge* about their presenting problems, thereby generating greater relational trust; (2) creating a secure *relationship* in which patients perceive their narratives to be recognized, marked and reflected back; (3) re-open the possibility for social communication (Fonagy, et al. 2019).

Despite the clinical and theoretical promise of ET as a concept, however, more empirical work is needed if we are to understand how to use it to inform clinical theory and practice. Trust already has an important role in the clinical thinking of Erikson (1953), Kohut (1982), and Bowlby (1969). Further studies need to emphasize what this picture can gain by focusing on *epistemic* trust, rather

than trust *simpliciter*. Even more crucially, researchers rarely address or attempt to measure *individual differences* in ET. These differences are necessarily at the heart of any clinical hypothesis concerning the construct ET, and work aiming to develop valid measures of them is warranted (Luyten, 2017; Nolte, 2017; but see Corriveau, et al., 2009).

In this paper, we present a validation study of a protocol for assessing ET: the Epistemic Trust Assessment (ETA, Schroeder-Pfeifer et al. 2018) . This paper begins by briefly outlining a definition of ET, its relationship with epistemic vigilance, and how they develop within early attachment relationships. We then describe in greater detail our theoretical assumptions for developing the ETA, the protocol for the validation study, and its results.

Epistemic trust

The ability to transmit and acquire cultural knowledge through the medium of overt interpersonal communication constitutes an important selective advantage for the human species (Csibra & Gergely, 2009). Humans are able to learn through reinforcement and social learning, but what seems unique to humans is the degree to which they are able to learn from overt and intentional communication addressed to them (Call & Tomasello). If this ability is to remain advantageous, however, humans need to remain vigilant against the risk of being misinformed (Sperber, et al. 2010). Learning from others is buttressed by a suite of cognitive mechanisms of *epistemic vigilance*, which determine the level of ET warranted in each context by taking into account different factors: e.g., the perceived reliability of the speaker, the consistency between what is communicated and previous knowledge of the addressee, etc. (Mercier & Sperber, 2019). Consequently, ET is conditional to the perceived reliability of communication and of the communicator.

ET and epistemic vigilance are thought to develop in early attachment relationships. Developmental research indicates that even at a very early age children do not treat all communication as equally trustworthy, and they take into account – among other things - evidence of the previous reliability of the communicator (Harris & Corriveau, 2011; Heyman, 2008; Koenig & Harris, 2007). An

experiment by Corriveau and Harris (2009) even suggests that child's attachment to the caregiver (i.e. his or her confidence in the caregiver as a source of protection) may also influence the tendency of the child to rely on her as an informant (i.e. epistemic trust).

In this experiment, children between 50 and 61 months were given pictures of animal hybrids (i.e. an animal comprised of two different animals, e.g., a fish and a squirrel; either in 50:50, or in 75:25 proportions). Asked to name these hybrids, children could enlist the help of their mother or a stranger, and then they were invited to endorse the claims of either of the two. Children's attachment to the caregiver had previously been tested when the children were 12 months old.

With respect to the 50:50 hybrids, the children were expected to be just as likely to label them as either one of the two animals comprising the hybrid (i.e. e.g., fish *or* squirrel). In this condition, secure and ambivalent children tended to agree with the label chosen by their mother rather than the label chosen by the stranger; this, however, was not the case with avoidant children, who picked the label chosen by the stranger just as often. On the other hand, with respect to 75:25 hybrids, children were expected to be *more* inclined to label them with the name of the animal that represented 75% of the hybrid. In this condition, however, mothers were instructed to always label the hybrid according to the name of the animal who represented 25% of the hybrid. In this second condition, secure and avoidant children - but not the ambivalent ones - tended to agree with the label chosen by the stranger, rather than the one chosen by their mother. To sum up, this experiment points to the possibility that early attachment may support the development of *individual differences* in epistemic trust. At one extreme, avoidant children may have learnt not to invest any special trust in the mother as an informant; at the other extreme, ambivalent children may be overly reliant on their mother's guidance. Secure children may occupy a well-judged middle ground – turning to a reliable informant when they need to, but relying on their own judgment whenever it is appropriate to do so.

Consistently with this experiment, Fonagy and his colleagues have recently proposed that secure attachment relationships may offer the ideal support for developing ET (see e.g., Fonagy & Allison, 2014). On the other hand, impairments in the capacity to exert epistemic vigilance and develop

epistemic trust may introduce biases in how we process interpersonally communicated information (Talia, Taubner, & Miller-Bottome, 2019), which may constitute a general vulnerability for psychopathology (Allison & Fonagy, 2016). Two meaningful impairments in ET can be anticipated. Some individuals may be unable to accept new or discordant information from others. Others might be predisposed to rely on others excessively for obtaining information. These ideas have strong resonance with psychoanalytic authors who have emphasized how mental disorders are associated with atypical strategies for learning from others (Bion, 1962; Lacan, 1958). They are also consistent with laboratory observations suggesting that severe psychopathology is associated with pervasive difficulties in establishing trust and acquiring information from others (e.g. Smeijers et al., 2017).

In Fonagy and colleagues' views, personality disorders can be seen as arising from a failure to establish ET in early attachment relationships and marked by persistent problems in communication underpinned by a lack of trust in interpersonally transmitted information (Allison & Fonagy, 2016; Fonagy et al., 2015; Fonagy & Allison, 2014). Accordingly, personality disorders would reflect communication disorders that prevent patients from updating their inner world based on information from the outside. Mental health, on the contrary, is underpinned by the capacity to exert appropriate vigilance in the face of possible deceit while maintaining general trust in interpersonally transmitted information. These views build on a widespread shift in emphasis in psychiatry from categorial diagnoses to generalized vulnerabilities for developing psychopathology (Selzam et al., 2018), in analogy with psychoanalytic models of neurosis (Lingiardi & McWilliams, 2017).

In this perspective, if atypical ET may be associated with mental disorders, re-establishing well-functioning epistemic vigilance may play an important role in facilitating therapeutic change (Fonagy & Allison, 2014). Fonagy and his colleagues have proposed three ways in which this process could occur. In the beginning of the therapeutic process, the common psychotherapy factors such as the therapist's proficiency and the theoretical framework of the therapy provide what Fonagy calls an "epistemic match", a foundation of ET in the therapeutic relationship to build upon. In the second phase or communication system, this foundation based on the structure of the therapy is extended to

information concerning relational and mental states within the therapeutic context, which rekindles the patient's capacity for mentalizing. In the last step, this process evolves beyond the therapeutic context and the general level of ET, required to learn from social experiences outside the therapeutic relationship, is restored.

Developing the Epistemic Trust Assessment

Because of the broad relevance of ET for psychopathology and clinical practice, a measure of ET is clearly needed. Such a measure could be used to understand better how ET is linked with psychopathology, and how it can be changed through psychotherapy. In the absence of a validated measure of ET, studies are beginning to approximate a measure of ET by resorting to self-reported participants' attachment (see Orme et al., 2019). However, our starting assumption in devising the ETA was that ET can only be measured implicitly and requires an experimental procedure or observer-based rating in order to be assessed. While the experience of trusting someone epistemically may be a conscious one, the process of vigilance that leads to epistemic trust is not. Vigilance mechanisms mainly comprise heuristics that are rapid, sub-personal, and "unconscious" (Sperber, et al. 2010).

Our initial intention was to adapt Corriveau's et al. procedure, described in the section above, to assess ET in adult participants. This aim, however, posed several challenges. The first challenge was to determine what sort of information could be supplied to our adult participants. In Corriveau et al.'s experiment, the use of encyclopedic information had the advantage of establishing "correct" and "incorrect" answers. However, encyclopedic information requires to control for participants' prior knowledge of the specific topic, which appears to be challenging with adult participants, who are likely to have different areas of expertise.

The second challenge was to create a procedure that would be relevant enough to the participants to trigger epistemic vigilance (Sperber et al., 2010). Experiments done by Gilbert and colleagues imply that new information presented to a subject will automatically be accepted as truthful if it is of no particular relevance to the subject himself or herself (Gilbert et al., 1990; Gilbert et al.,

1993). Hence, in order to assess ET, one needs to be ensured that the information presented to participants is relevant to them. While it is comparatively simple to establish relevance for information in an experimental context with young children who, for the most part, do not have consolidated interests or knowledge, it is more complex with adolescents or adults.

To address both challenges, we included in our procedure the Trier Social Stress Test for Groups (TSST-G, Dawans et al., 2011) as a primer to our experiment. The TSST-G is a standardized procedure that reliably induces social stress utilizing a mock job-interview conducted by experts. The mock job-interview, or more precisely, how the participants performed in the interview, provided a standardized information that could tap both into declarative and non-declarative content. We hypothesized that information about performance in a job-interview situation would be inherently relevant to the participants as it would likely be useful for their future career. Additionally, research suggests that individuals experiencing stress actively take in more information from external sources to counteract the uncertainty arising from a stressful situation (Driskell & Salas, 1991; Eysenck et al., 2007; Starcke & Brand, 2012).

After establishing the context of the information transmitted in the experiment, another challenge occurred with regard to the content of the information transmitted. Trusting the relevance and generalizability of a given piece of information relies on a number of different factors, which must be taken into account by the addressee. Namely, the addressee, in this case the participant, will establish epistemic trust based on their previous experience of the communicator's reliability, the communicator's competence on the specific topic, the congruence between what the communicator says and the addressee's previous knowledge of the topic, and any arguments or evidence provided in support by the communicator (Mercier & Sperber, 2017). Since the participant, does not know the communicator, aka the committee evaluating the participant's performance in the TSST-G, the participants only have their first impression to work with. Consequently, we included different types of content in the information about the participants TSST-G performance that may be associated with

different levels of epistemic trust. In order from the lowest to the highest inherent level of certainty, we divided our feedback content into information about physiological, relational, and mental states.

Individuals differ in their knowledge about one's physiological features such as heart rate or blood pressure (Shah et al., 2017; Zamariola et al., 2018). Consequently, we hypothesized that this category of information should be prone to be influenced by outside information. This should be especially true if the committee of the TSST-G, in addition to their expert status, have technological aides to draw upon, such as heart-rate monitors, which makes them trustworthy informants for this category of information. Since the TSST-G usually involves measurements of either cortisol, heart rate, or both, this was easily accomplished and supposed to make the information seem even more reliable. While this type of information might usually not be very relevant outside of context (e.g. the information being critical for one's health), the job-interview in the TSST-G was considered to provide sufficient relevance.

With regard to knowledge about relational states, all partners in a social encounter have individual, shared intrapersonal and interpersonal subjective information about their social encounter (Talia, Miller-Bottome, Wyner, Lilliengren, & Bate, 2019). While each partner certainly has an opinion about how he/ she is perceived from the outside, they can never be sure. Thus, relational states should be characterized by a moderate amount of inherent certainty and be partly open to outside information.

With regard to mental states, it is generally assumed in communication that we have privileged access to our own mental states, such as thoughts, feelings, and beliefs (Santarelli & Talia, 2017). Statements about one's mental states should therefore be characterized by the highest inherent certainty. They should thus be the least open to outside information, if said information is incongruent with one's own perception. Consequently, the committee should be an untrustworthy informant for information in this category, if the information is incongruent with the participants own opinion.

Summarizing, the ETA is an experimental procedure in which the participants are first asked to provide information about themselves, are then provided feedback on that information by an expert

committee, and third are asked to re-rate that same information about themselves. The individuals' ET is operationalized as the differences between the initial assessment (pre assessment) and the assessment after having received feedback (post assessment).

The aim of this study is to explore whether or not different patterns emerge both in initial certainty and in the change of certainty after being presented with the feedback. Accordingly, two hypotheses are tested:

1. Initial certainty differences – Building on the opaqueness of physiological states in contrast to the privileged access to one's own mental states, we hypothesized that there would be significant differences in initial certainties according to these categories. Specifically, we hypothesized that physiological states would have the lowest initial certainty, followed by relational states, and mental states with the highest initial certainty.

2. Change in certainty differences – Building on the first hypothesis, we assumed that the change in certainty from before receiving feedback to afterwards would significantly differ depending on the category. Because ET is a balanced construct, which means we only take in information from the outside where it makes sense instead of taking in all or none of the information, we hypothesized that the change in certainty pre feedback to post would be the largest for physiological states, followed by relational states, with mental states showing the least change due to receiving the feedback.

In addition to these hypotheses, we investigated whether different patterns of the core study variables emerged in the present non-clinical sample. Since ET is a construct involved in everyday communication and learning, it seems feasible that different profiles of ET might emerge in this healthy sample. In this, we do not expect extreme divergence from epistemic vigilance, but rather more subtle differences. For example, some individuals might have higher initial certainties while still adjusting as much as others in response to the feedback in the experiment. This would indicate a subtype, who might appear assertive in communication but is still able to adjust his or her views based on information from the environment in contrast to a pathological type, who appears assertive and is unable to change his views.

Methods

Recruitment and Participants

Participants were students from the University of Heidelberg (Germany), who voluntarily signed up to participate in studies via an online study participation platform. The platform had 2,424 registered students at the time of recruitment. Inclusion criteria were age above 18, ability to provide informed consent, and fluency in the German language. A total of *N* = 62 students took part in the study, of these 20% studied psychology and 5% studied medicine, 73% studied other majors and 2% did not actively pursue a major but were still inscribed. The mean age was *M* = 25.21 (range from 19 to 61) and 69% of participants were women. One participant dropped out during the interview part of the TSST-G due to hypotension. The study was conducted with ethical approval from the ethics committee of the Medical Faculty of the University of Heidelberg, Germany (reference number: S-272/2017).

Measures

Epistemic Trust Assessment (ETA)

The ETA is an experimental procedure that aims to measure ET. It includes four different phases. In the *first phase*, the participant undergoes a mock job-interview according to a procedure commonly used in the TSST-G. The TSST-G (Dawans et al., 2011) is a standard paradigm for the reliable induction of moderate social stress and the group version of the original paradigm (Kirschbaum et al., 1993) and was chosen for economic reasons as it allows for the assessment of up to six participants at a time in front of a committee of two people in doctors coats. The participants first undergo a mock job-interview. They are instructed to prepare themselves for a two-minute job-interview (6 min.). Participants can take notes during preparation, but are not allowed to use them during the interview. Afterwards the participants are called at random to deliver their interview one after another. During

the interview the committee interrupts the participant with statements like “Thank you, that is less interesting to us. Could you describe your problem-solving behaviour?”. After each participant of the group has been interviewed for two minutes, they are instructed to count backwards from different four digit numbers in increments of 13. The committee interrupts the participants if they are incorrect or they are prompted to go faster if they are correct. Application of the TSST-G has been linked to an increase in the activation of the hypothalamic–pituitary–adrenal stress system, as well as eliciting an autonomic stress response (Boesch et al., 2014; Leder et al., 2013).

In the *second phase*, the participants are asked to rate themselves on how they performed during the TSST-G by individually filling out a self-report questionnaire on a computer. The questionnaire includes statements that inquiry about physiological states (i.e. “Was your pulse below or above 96 during the experiment?”), relationship aspects (i.e. “Do you think you came across as friendly during your interview?”), and mental states (i.e. “Were you anxious during the experiment?”) during the interview. We chose four statements per category in an effort to minimize effort for the participants while enabling just-identified confirmatory factor analysis.

In the third phase (feedback), the participants are presented with a computer-generated feedback, which is presented as if it was provided by the expert committee. The feedback is programmed to be congruent with the participant’s valence rating in exactly half of the questions (chosen at random among each of the three groups), in order to avoid a bias on over- or under-agreement.

In the fourth phase (post-feedback re-assessment), the participants are asked to re-rate all the items from the first phase, taking into account the feedback from the expert committee (third phase). All items are rated on a 0 to 100 scale ranging from 0% certainty to 100% certainty in the first phase and from 0% agreement with the committee to 100% agreement with the committee in the third phase.

The questionnaire has three subscales: physiological states (P), relational states (R), and mental states (S), including four items each. For the physiological and relational subscales, a change score is

computed by subtracting the mean score of the post-feedback items from the mean score of the self-assessment items. For the mental states subscale a score is computed by subtracting the post-feedback item mean score of the two items where the committee did not agree with the participant from the self-assessment mean score. This is done so that a maximum score in the mental states subscale represents not accepting incongruent information about mental states. Finally, a total ET score can be computed from the ETA with the following formula:

$$ETScore = (ChangeP) + (100 - |ChangeR|) + (-ChangeS)$$

The score has a range of 300, indicating adaptive ET to -200 indicating maladaptive ET. The ETA was operationalized using RShiny (Chang et al., 2017).

Short Scale Social Desirability-Gamma

The Short Scale Social Desirability-Gamma (Kurzskala Soziale Erwünschtheit-Gamma; KSE-G; Kemper et al., 2012) is an economic measure for the assessment of social desirable behavior (Paulhus, 2015). The KSE-G consists of three items each loading on two scales, exaggeration of positive qualities and understating of negative qualities. The items describing social behavior (i.e. “When in an argument, I always stay factual and objective”) on a 5-point Likert scale ranging from “does not apply at all” to “applies fully”. According to the authors, the instrument exhibits satisfactory internal consistency and high factorial and content validity (Kemper et al., 2012).

Inventory of Personality Organization

The 16-item version of the Inventory of Personality Organization (IPO-16, Zimmermann et al., 2013) is a unidimensional self-report measure assessing personality functioning based on Kernberg’s model of personality organization (Kernberg, 1984). The IPO-16 assesses identity diffusion, primitive psychological defences and reality testing and is rated on a 5-point Likert scale ranging from “never applies” to “always applies”. The authors supply cut-off values based on ROC-analyses using

established personality functioning measures and report good internal constancy (α = .85), good discriminant, and convergent validity (Zimmermann et al., 2013).

The ETA, KSE-G and IPO-16 were administered on a computer screen with an R-Shiny adaptation.

Procedure

In order to control for possible outliers due to psychopathology we employed a brief screening measure of personality functioning (Zimmermann et al., 2013). Additionally, since the participants were university students, we asked whether or not they majored in psychology or medicine, in order to control for this in the statistical analysis. Psychology students might be familiar with psychological tests like the TSST (and might be aware of the mock nature of the committee), which in turn might undermine the relevance of the paradigm for these students. Medical students might also realize that the feedback on physiological states is not authentic, as we only built in a three-minute break during which the data from the heartrate monitors were supposedly evaluated.

Students, who had signed up for participation in the study via the university's online platform, were sent an email with information about the study and an outline of the experimental procedure. Upon arrival at the study site, participants were provided detailed information about the type of data assessed in the experiment, the procedure of the assessment, their benefits and risks in participating in the study, as well as contacts for further information and assurance that they could discontinue participation at any point in time.

The underlying aim of the study was obscured and instead the study's aim was described as exploring the relationship between stress and personality, as well as physiological attributes. After providing informed consent, the participants were asked to complete both the IPO-16 and KSE-G before undergoing the TSST-G as per protocol (Dawans et al., 2011). After the TSST-G, the ETA was administered. Finally, the participants were debriefed about the aim of the study and compensated with 10€.

Statistical Analysis

An a priori power analysis was computed, for a medium effect of $f^2 = .25$, an alpha of $\alpha = .05$ and a power of $\beta = .90$, with a resulting sample size of $n = 54$. We assumed a drop-out rate of 10% thus $n = 60$ participants were to be recruited. Details of the power analysis for this study can be found elsewhere (Schroeder-Pfeifer et al. 2018).

In order to answer the question, if the mean certainty ratings pre and post-feedback per category were highest for the physiological and lowest for the internal states, a two-sided ANCOVA was conducted. As described above, we controlled for a major in medicine or psychology. Additionally, we controlled for sex in case of any effects of sex on certainty ratings or acceptance of feedback. We chose a forced choice answer format in the app questionnaires to achieve complete data for all participants with no missing values.

Visual inspection of a Quantile-Quantile plot of the residual quantiles against the theoretical quantiles did not indicate non-normality of the error for any analysis. Visual analysis of the fitted values plotted against the residual values did not indicate heteroscedasticity for any analysis. No multivariate outliers were found as the largest within-cell Mahalanobis' distance (47.97) was smaller than the χ^2 critical value of 54.7 (33, 61, $p < .001$).

A hierarchical cluster analysis was conducted in an effort to extract patterns of different groups of participants with regard to the core study variables. In this, an agglomerative (bottom-up) approach utilizing complete linkage was chosen. As there are no significant outliers in this sample which might discourage the use of complete linkage, we utilized this algorithm as it avoids chaining problems typically encountered for single linkage approaches. Euclidean distance was used to compute the dissimilarity matrix. R (version 3.5.2, R Development Core Team, 2008) was used for all statistical analyses. The R code and anonymized data to recreate the analysis in this paper are available in a reproducible cloud setting (https://doi.org/10.24433/CO.1275451.v2).

Results

A table showing the descriptive statistics of age, gender, ET, personality functioning and social desirability can be found in the appendix.

Initial certainty differences

Table 1 reports estimates of the ANCOVA predicting initial level of certainty from category of statement, with gender and psychology/ medicine major as covariates. The lower half of the table displays the results of the Tukey post-hoc test. All differences between the categories of statements were significant in the expected direction, indicating that physiological states had on average the lowest initial certainty, followed by relational states, with self-states having the highest initial certainty. This is congruent to what we stated in the first hypothesis, both in terms of differences between the categories as well as the order of categories according to initial certainty.

Table 1

Estimates of ANCOVA predicting mean level of initial certainty

Dependent variable: Mean level initial certainty

Variable	*df*	*F* value	*p*
Category	2	40.29	**<.001**
Med. /Psy. Major	1	0.60	.411
Gender	1	0.33	.566
Tukey HSD	diff	95% CI	adj. *p*
Initial R vs Initial P	7.29	0.79 – 13.79	**.024**
Initial S vs Initial P	24.08	17.58 – 30.58	**<.001**
Initial S vs Initial R	16.79	10.29 – 23.29	**<.001**

Note. CI = confidence interval; Category = Item Category of the ETA; Med. /Psy. Major = did the participant major in either medicine or psychology; Initial P = Initial certainty on the ETA physiological scale; Initial R = Initial certainty on the ETA relational scale;
Initial S = Initial certainty on the ETA self-states scale.

Changes in certainty

Table 2 reports estimates of the ANCOVA predicting change in certainty from category of statement, with gender and psychology major as covariates. Neither of the covariates had a significant association with the mean level of initial certainty, leaving category of statement as the only significant predictor. The result of the Tukey post-hoc tests for the ANCOVA displayed in the lower half of the table shows the differences in mean certainty between all three categories of statements as significant. These results are in line with our second hypothesis, both in terms of differences between the categories as well as the order of categories according to change in certainty.

Table 2

Estimates of ANCOVA predicting mean level of change in certainty

Dependent variable: Mean level of change in certainty

Variable	*df*	*F* value	*p*
Category	2	104.91	**<.001**
Med. /Psy. Major	1	0.04	.848
Gender	1	1.28	.260
Tukey HSD	diff	95% CI	adj. *p*
Change R vs Change P	-18.25	-28.61 - -7.90	**<.001**
Change S vs Change P	-61.79	-72.14 - -51.43	**<.001**
Change R vs Change P	-43.53	-53.90 - -33.17	**<.001**

Note. CI = confidence interval; Category = Item Category of the ETA; Med. /Psy. Major = did the participant major in either medicine or psychology; Initial P = Initial certainty on the ETA physiological scale; Initial R = Initial certainty on the ETA relational scale;
Initial S = Initial certainty on the ETA self-states scale.

Cluster analysis

Three major clusters were identified in the dendogram (see appendix) and by the elbow criterion in the scree plot. Individuals from the third cluster were on average more prone to change in the physiological and relational categories than those from the other two clusters (table 2 in the appendix). In addition, the initial certainty values for the physiological and the relational category was

lower than in the other clusters as well. While the standardized IPO-16 sum scores were significantly higher in cluster three opposed to cluster two, this difference was in no way clinically significant because of the low variance in IPO-16 sum scores in the entire sample. Participants in the second cluster were much more prone to change their certainty for self-states than participants in both of the other clusters. Additionally, they also had much lower initial certainties in self-states than participants in either of the other clusters. Participants in cluster one were characterized by the lowest change in certainty across all three categories, as well as having the highest initial certainties.

Discussion

The aim of this study was to devise and validate a procedure for assessing ET in adult participants. Consistent with our hypothesis, we found a significant difference in the mean level of inherent certainty per category of statement (low certainty for physiological, moderate for relational, high for self-states) as well as in the change in certainty per category of statement (high changes for physiological, moderate for relational, low for self-states). The agglomerative cluster analysis found three clusters that may be considered to represent an overly vigilant, naïve, and open subtype. These clusters differ significantly on all variables but age and the exaggerating positive qualities scale of the KSE-G.

The cluster analysis hinted at three distinct subgroups of participants: an overly vigilant subgroup, a naïve/uncertain subgroup, and an open subgroup. The overly vigilant subgroup was characterized by relatively high initial certainties as well as little change post feedback. The naïve/ uncertain group was characterized by low initial certainty in self-states, as well as high change in certainty in the self-states-category. The open subgroup had low initial certainties in physiological as well as relational states, as well as high change in the physiological and relational categories. The findings closely resemble the theoretically described clusters for ET (Fonagy & Allison, 2014).

Our results suggest that, as hypothesized, subjects tend to revise their previous beliefs about the self on the basis of interpersonally communicated information, and that they do so with greater

or lesser ease according to the perceived reliability of the informant about the topics discussed. This conclusion is especially interesting from the vantage point of improving our understanding how psychotherapy works. Psychotherapy typically concerns itself with facilitating change in patients' assumptions through interpersonal communication with the therapist. As emphasized by our results, however, not all communication is equally likely to lead to such a change. In our study, subjects were more likely to change their opinion about their physiological states (which the committee was presented as expert of) than about their own mental states. Similarly, the therapist would by most patients not be considered an "expert" of one of his or her patient's ongoing mental states. Therapists' communication may thus be perceived as less reliable when they provide overly certain feedback about this subject, and perhaps especially so if the relationship with the informant is characterized by moderate stress (as in our procedure). Thus, trust establishing feedback should maybe aim at feedback about the relationship instead while disclosing the therapist's own subjective experience. This conclusion seems highly relevant to psychotherapy, where such discussions are especially frequent and relevant. However, this may be different in distinct clinical groups and remains a subject of future studies.

Limitations and future directions

A possible effect known from research on metacognitive phenomena that may have interfered with our ET experiment is the so called hypercorrection effect (e.g. Butterfield & Metcalfe, 2001; Metcalfe & Finn, 2011). This effect describes a tendency to more easily correct apparently wrong statements that were of high prior certainty as opposed to low prior certainty. This might lead to participants overcorrecting statements with high inherent certainty, such as from the relational and mental states category. However, this effect has not yet been thoroughly examined for non-declarative information, and research suggests that participants have to be relatively sure that the alternative statement provided to them is correct feedback (Metcalfe & Finn, 2011). In the face of non-declarative information like the feedback provided by the committee in this study, it may be possible that this effect applies to the category with physiological information, since both relational and mental state

information is inherently subjective and can thus never be entirely “correct”. Furthermore, since our sample only entailed healthy university students with little to no personality disorder traits, the paradigm may be considered to lack external validity. In our sample, four participants scored above the lower IPO-16 cut-off proposed by Zimmermann et al. (2013) and only one participant scored above the higher cut-off of 2.38. The mean in this sample was markedly below the cut-off at *M* = 1.3 with little variance (*sd* = 0.43). As such it is also not surprising that the derived scores of the ETA were not correlated with the IPO-16 scores.

In order to further develop our experimental paradigm, future studies may consider using different items and examine more diverse samples i.e. a clinical sample, a personality disorder sample, or a more representative population sample not only sampling from students. Additionally, little is known with regard to how ET relates to cultural differences. Following, replication efforts of this study in culturally diverse samples are needed. In this study, we used the TSST-G in order to induce stress. However, to achieve this effect, it could also be tested to use another, more economic test, e.g. the socially evaluated cold-pressor test (Schwabe & Wolf, 2010). While the procedure would have to be slightly altered to include a more marked element of social evaluation, it may considerably improve the viability of the paradigm.

Future studies should use the ETA to compare groups with low personality functioning (including patients diagnosed with borderline personality disorder or antisocial personality disorder) or mental disorders with higher functioning and healthy individuals. Applying the ETA to these populations is urgently needed if we are to find empirical evidence for the hypothesis that ETA constitutes a broad vulnerability for developing psychopathology. In order to thoroughly examine the ETA’s relation to measures of personality disorder, a clinical sample of patients with personality disorder or a mixed sample would be needed. would be the ideal to test the ET related clinical theory on those groups of patients (Fonagy et al., 2015).

In the same vein, another axis which future studies using the ETA could explore to gain more insight into the mechanisms of ET in general and in psychotherapy specifically is the use of ostensive

cues. Systematically varying the use of ostensive cues on part of the committee especially in groups of participants with low ET, might shed light on whether or not it is possible to overcome initially low ET by establishing relevance via ostensive cues.

Conclusion

This study provides a first theoretically grounded operationalization of individual differences in ET in adults. A distinct advantage of the concept of ET is that it appears to be compatible with the theoretical perspectives of many therapy schools, as it focuses on a general requirement of human learning and change. At the same time, the concept has its roots in psychoanalytic theory. Consistent with psychoanalytic approaches, recent thinking on ET emphasizes how individual differences in early development may come to create a basic vulnerability for psychopathology, expressed through communication at an intrapsychic and an interpersonal level. Future research should expand what we know about such differences in psychopathology and in psychotherapy, and investigate how to tailor insight- and relation-focused therapeutic work to these patients' epistemic needs.

Declaration of interest statement

None of the authors has any conflict of interest to declare.

References

Alexander, F., & French, T. M. (1946). The corrective emotional experience. *Psychoanalytic therapy: Principles and application*.

Allison, E., & Fonagy, P. (2016). When is truth relevant? *The Psychoanalytic Quarterly*, *85* (2), 275–303. https://doi.org/10.1002/psaq.12074

Bergstrom, B., Moehlmann, B., & Boyer, P. (2006). Extending the testimony problem: Evaluating the truth, scope, and source of cultural information. *Child Development*, *77* (3), 531–538. https://doi.org/10.1111/j.1467-8624.2006.00888.x

Bion, W. R. (1962): Learning from experience. London: Maresfield Library.

Boesch, M., Sefidan, S., Ehlert, U., Annen, H., Wyss, T., Steptoe, A., & La Marca, R. (2014). Mood and autonomic responses to repeated exposure to the Trier Social Stress Test for Groups (TSST-G). *Psychoneuroendocrinology*, *43*, 41–51. https://doi.org/10.1016/j.psyneuen.2014.02.003

Boston Change Process Study Group. (2005). The "something more" than interpretation revisited: Sloppiness and co-creativity in the psychoanalytic encounter. *Journal of the American Psychoanalytic Association*, *53*(3), 693-729.Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27* (6), 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491

Bowlby, J. (1969): Attachment and loss. Penguin books (The international psycho-analytical library, no. 79; 95; 109).

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187–192.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2017). *Shiny: Web Application Framework for R* [R package version 1.0.3.]. https://CRAN.R-project.org/package=shiny

Corriveau, K. H., & Harris, P. L. (2009). Choosing your informant: Weighing familiarity and recent accuracy. *Developmental Science*, *12* (3), 426–437. https://doi.org/10.1111/j.1467-7687.2008.00792.x

Corriveau, K. H., Harris, P. L., Meins, E., Fernyhough, C., Arnott, B., Elliott, L., Liddle, B., Hearn, A., Vittorini, L., & Rosnay, M. de (2009). Young children's trust in their mother's claims: Longitudinal links with attachment security in infancy. *Child Development*, *80* (3), 750–761. https://doi.org/10.1111/j.1467-8624.2009.01295.x

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13* (4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Dawans, B. von, Kirschbaum, C., & Heinrichs, M. (2011). The Trier Social Stress Test for Groups (TSST-G): A new research tool for controlled simultaneous social stress exposure in a group format. *Psychoneuroendocrinology*, *36* (4), 514–522. https://doi.org/10.1016/j.psyneuen.2010.08.004

Driskell, J. E., & Salas, E. (1991). Group decision making under stress. *Journal of Applied Psychology*, *76* (3), 473–478. https://doi.org/10.1037/0021-9010.76.3.473

Egyed, K., Kiraly, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24* (7), 1348–1353. https://doi.org/10.1177/0956797612471952

Fonagy, P., & Allison, E. (2014). The role of mentalizing and epistemic trust in the therapeutic relationship. *Psychotherapy*, *51* (3), 372–380. https://doi.org/10.1037/a0036505

Fonagy, P., & Campbell, C. (2015). Bad blood revisited: Attachment and psychoanalysis, 2015. *British Journal of Psychotherapy*, *31*(2), 229-250.

Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic Petrification and the Restoration of Epistemic Trust: A New Conceptualization of Borderline Personality Disorder and Its Psychosocial Treatment. *Journal of Personality Disorders*, *29* (5), 575–609. https://doi.org/10.1521/pedi.2015.29.5.575

Freud, S. (1961). The ego and the id. In *The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIX (1923-1925): The Ego and the Id and Other Works* (pp. 1-66).

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59* (4), 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, *65* (2), 221–233. https://doi.org/10.1037/0022-3514.65.2.221

Hagá, S., & Olson, K. R. (2017). Knowing-it-all but still learning: Perceptions of one's own knowledge and belief revision. *Developmental Psychology*, *53* (12), 2319–2332. https://doi.org/10.1037/dev0000433

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *366* (1567), 1179–1187. https://doi.org/10.1098/rstb.2010.0321

Heyman, G. D. (2008). Children's Critical Thinking When Learning From Others. *Current Directions in Psychological Science*, *17* (5), 344–347. https://doi.org/10.1111/j.1467-8721.2008.00603.x

Jones, E. E. (2000). *Therapeutic action: A guide to psychoanalytic therapy*. Rowman & Littlefield.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A., & Rammstedt, B. (2012). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G) [A short scale for the assessment of the Gamma-Factor of social desirability: the short scale for social desirability-gamma]*. *GESIS*, http://nbn-resolving.de/urn:nbn:de:0168-ssoar-339589

Kernberg, O. F. (1984). *Severe personality disorders: Psychotherapeutic strategies*. Yale University Press. http://www.jstor.org/stable/10.2307/j.ctt32bf53 https://doi.org/10.2307/j.ctt32bf53

Kernberg, O. F., Diamond, D., Yeomans, F. E., Clarkin, J. F., & Levy, K. N. (2008). Mentalization and attachment in borderline patients in transference focused psychotherapy. In E. Jurist, A. Slade, & S. Bergner (Eds.), Mind to mind: Infant research, neuroscience and psychoanalysis (pp. 167–201). New York, NY: Other Press.)

Kirschbaum, C., Pirke, K.‐M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' - A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology* (28), 76–81. https://doi.org/10.1159/000119004

Koenig, M. A., & Harris, P. L. (2007). The Basis of Epistemic Trust: Reliable Testimony or Reliable Sources? *Episteme: A Journal of Social Epistemology*, *4* (3), 264–284. https://doi.org/10.1353/epi.0.0017

Lacan, Jacques (1958): Écrits. New York, London: W.W. Norton.

Lingiardi, V., & McWilliams, N. (Eds.). (2017).Psychodynamic Diagnostic Manual(2nd ed.). New York, NY: Guilford Press.

Leder, J., Hausser, J. A., & Mojzisch, A. (2013). Stress and strategic decision-making in the beauty contest game. *Psychoneuroendocrinology*, *38* (9), 1503–1511. https://doi.org/10.1016/j.psyneuen.2012.12.016

Luyten, P. (2017). *Epistemic trust and BPD: An experimental approach.* ISSPD, Heidelberg.

Mercier, Hugo; Sperber, Dan (2019): The Enigma of Reason: Harvard University Press.

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: Did they know it all along? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37* (2), 437–448. https://doi.org/10.1037/a0021962

Nolte, T. (2017). *Epistemic trust in adolescents and BPD patients: Two experimental approximations.* ISSPD, Heidelberg.

Orme, W.; Bowersox, L.; Vanwoerden, S.; Fonagy, P.; Sharp, C. (2019): The relation between epistemic trust and borderline pathology in an adolescent inpatient sample. In *Borderline personality disorder and emotion dysregulation* 6, p. 13. DOI: 10.1186/s40479-019-0110-7.

Paulhus. (2015). Socially desirable responding: The evolution of a construct. In H. Braun, D.N., Jackson, & D. E. Wiley (Eds.), *The Role of Constructs in Psychological and Educational Measurement.* Routledge.

R Development Core Team. (2008). *A language and environment for statistical computing.*

Schroder-Pfeifer, P., Talia, A., Volkert, J., & Taubner, S. (2018). Developing an assessment of epistemic trust: a research protocol. *Research in Psychotherapy: Psychopathology, Process, and Outcome*, *21*(3).Santarelli, M., & Talia, A. (2017). A Pragmatist Perspective On Self-State Knowledge In The Therapy Context. *Pragmatism Today*, 8(1), 133-145.

Schwabe, L., & Wolf, O. T. (2010). Socially evaluated cold pressor stress after instrumental learning favors habits over goal-directed action. *Psychoneuroendocrinology*, *35* (7), 977–986. https://doi.org/10.1016/j.psyneuen.2009.12.010

Selzam, S., Coleman, J. R., Caspi, A., Moffitt, T. E., & Plomin, R. (2018). A polygenic p factor for major psychiatric disorders. *Translational psychiatry*, *8*(1), 1-9.

Smeijers, D., Rinck, M., Bulten, E., van den Heuvel, T., & Verkes, R. J. (2017). Generalized hostile interpretation bias regarding facial expressions: Characteristic of pathological aggressive behavior. *Aggressive behavior*, *43*(4), 386-397.

Sperber, D. (2001). An Evolutionary Perspective on Testimony and Argumentation. *Philosophical Topics*, *29* (1), 401–413. https://doi.org/10.5840/philtopics2001291/215

Sperber, D., Clement, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language*, *25* (4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Sperber, D., & Wilson, D. (2001). *Relevance: Communication and cognition* (2nd ed.). Blackwell Publishers.

Strachey, J. (1934). The nature of the therapeutic action of psychoanalysis. *Classics in psychoanalytic technique*, 361-378.

Talia, A., Miller-Bottome, M., Wyner, R., Lilliengren, P., & Bate, J. (2019). Patients' Adult Attachment Interview classification and their experience of the therapeutic relationship: are they associated?. Research in Psychotherapy: Psychopathology, Process and Outcome, 22(2). 10.4081/ripppo.2019.361

Talia, A., Taubner, S., & Miller-Bottome, M. (2019). Advances in research on attachment-related psychotherapy processes: seven teaching points for trainees and supervisors. Research in Psychotherapy: Psychopathology, Process and Outcome, 22(3). https://doi.org/10.4081/ripppo.2019.405

Wilson, D., & Sperber, D. (2012). *Meaning and relevance*. Cambridge University Press.

Zimmermann, J., Benecke, C., Hörz, S., Rentrop, M., Peham, D., Bock, A., Wallner, T., Schauenburg, H., Frommer, J., Huber, D., Clarkin, J. F., & Dammann, G. (2013). Validierung einer deutschsprachigen 16-Item-Version des Inventars der Persönlichkeitsorganisation (IPO-16). *Diagnostica*, *59* (1), 3–16. https://doi.org/10.1026/0012-1924/a000076

Appendix

Table 1

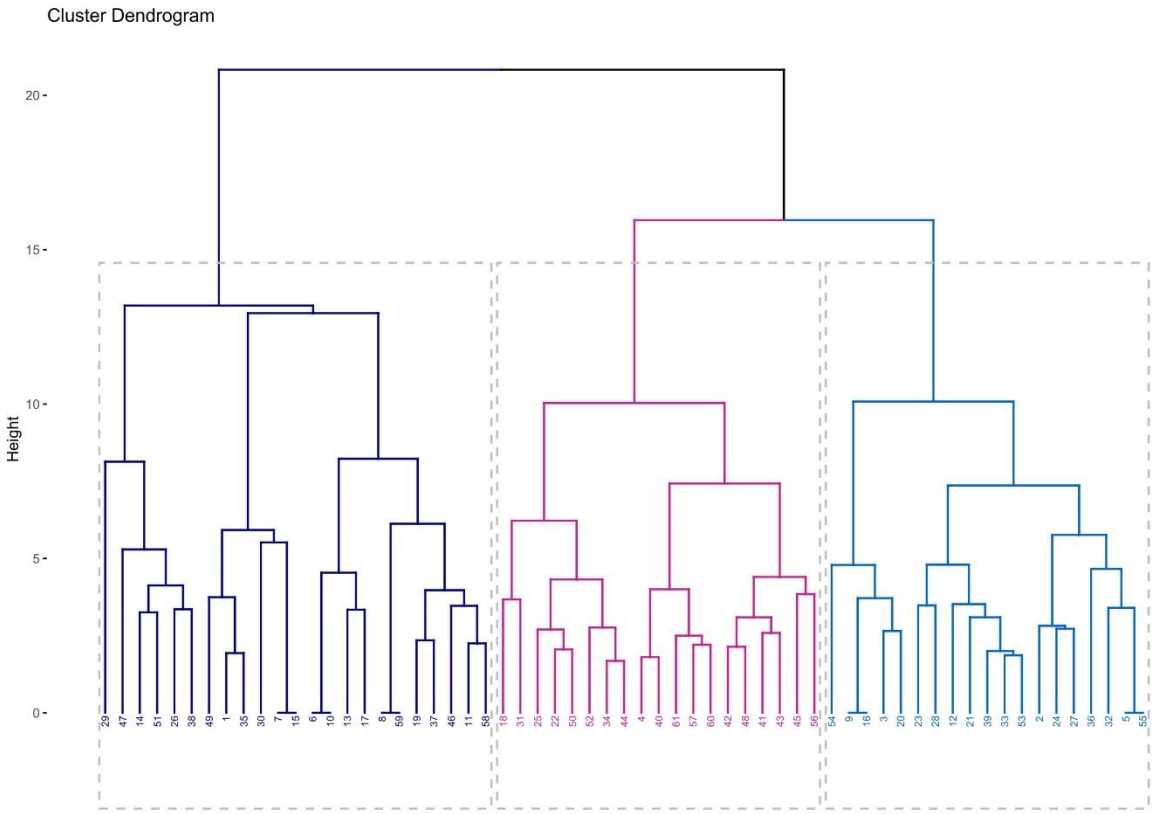
Descriptive statistics of core study variables

Variable	*M*/%	*SD*	Min	Max	Skew	Kurtosis
Age	25.21	8.84	19.00	61.00	2.68	6.76
Female gender	69%	-	-	-	-	-
ETQ:						
Initial certainty physiological states	56.12	19.87	0.00	90.00	-0.61	0.09
Initial certainty relational states	63.41	10.95	38.75	94.00	-0.02	0.10
Initial certainty self-states	80.2	13.18	50.00	100.00	-0.33	-0.97
Change in certainty physiological states	13.89	27.44	-53.5	73.50	0.07	-0.31
Change in certainty relational states	-4.36	16.28	-47.25	34.75	-0.40	0.09
Change in certainty self-states	-47.89	27.07	-97.5	13.25	0.21	0.21
ET Score	125.11	32.45	13.75	206.75	-0.39	1.19
ET Naïveity	-14.58	44.56	-118.75	68.75	-0.28	-0.47
Personality functioning:						
IPO-16 sum score	1.30	0.43	0.50	2.50	0.41	0.06
Social desirability (KSE-G):						
Minimizing negative qualities	2.63	0.89	0.00	4.00	-0.74	0.36
Exaggerating positive qualities	2.45	0.61	1.00	3.67	-0.67	0.14

Table 2

Means per cluster as well as MANOVA and post-hoc test significance

Variable	Cluster 1: "Overly Vigilant" *n* = 23	Cluster 2: "Naïve/ Uncertain" *n* = 19	Cluster 3: "Adaptive" *n* = 19			
	z-scores/ percentage	*z*-scores/ percentage	*z*-scores/ percentage	*F(df) / x²(df)*	*p*	Post Test
Age	0.39	-0.28	-0.20	2.92(2)	.057	-
Change certainty P	-0.42	0.02	0.54	6.13(2)	**.007**	3-1
Change certainty R	-0.51	-0.09	0.71	4.47(2)	**<.001**	3-1; 3-2
Change certainty S	-0.69	0.94	-0.10	31.50(2)	**<.001**	2-1; 3-1; 3-2
Initial certainty P	0.43	0.17	-0.70	6.37(2)	**<.001**	3-1; 3-2
Initial certainty R	0.64	0.15	-0.93	12.78(2)	**<.001**	3-1; 3-2
Initial certainty S	0.36	-0.74	0.31	14.52(2)	**<.001**	2-1; 3-2
IPO16 Sum	0.13	-0.53	0.37	6.81(2)	**.013**	2-1; 3-2
Female Sex	47.83%	73.68%	89.47%	8.20 (2)	**.013**	3-1
Ex. Pos. qualities	0.00	-0.02	0.01	1.911 (2)	.996	-
Min. neg. qualities	-0.48	0.30	0.28	1.40(2)	**.012**	2-1; 3-1


Cluster Dendrogram
Height
20
15
10
5
0
29 47 14 51 26 38 49 1 35 30 7 15 6 10 13 17 8 59 19 37 46 11 58
18 31 25 22 50 52 34 44 4 40 61 57 60 42 48 41 43 45 55
54 9 16 3 20 23 28 12 21 39 33 53 2 24 27 36 32 5 55

Study 4

V. Zettl, M., Back, S.N., Taubner, S., & **Schröder-Pfeifer, P.** (Under review) Assessing Personality Functioning and Maladaptive Traits in Young Adults: A Machine Learning Approach. *Journal of Personality Disorders.* IF: 2.440

Declaration of author contributions: Zettl, M.: Conceptualization, , Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Project Administration, Funding Acquisition **Back, S.N.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing **Taubner, S.:** Writing – Review & Editing, Supervision **Schröder-Pfeifer, P.:** Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – Original Draft, Writing – Review & Editing

Title: Assessing Personality Functioning and Maladaptive Traits in Young Adults: A Machine Learning Approach

Max Zettl

Department for Psychosocial Prevention, University of Heidelberg

Sarah N. Back,

Department of Psychology, Ludwig Maximilian University Munich

Svenja Taubner

Department for Psychosocial Prevention, University of Heidelberg

Paul Schröder-Pfeifer

Department for Psychosocial Prevention, University of Heidelberg

Author Note

Correspondence concerning this article should be addressed to Max Zettl, Departement for Psychosocial Prevention, University of Heidelberg, Bergheimer Str 54, 69115 Heidelberg, Germany, Phone: 0049 6221 56 410, Fax: 0049 6221 56 4702 E-Mail: Max.Zettl@med.uni-heidelberg.de

Abstract

The assessment of personality disorders (PDs) is complex and often neglected. Moreover, current theories of PDs lack scope, comprehensiveness, and empirical support. However, the emergence of artificial intelligence approaches and empirically derived PD models has the potential to guide clinical assessment and advance personality theory. In this study, we apply machine learning (ML) to PD assessment in a sample of young adults. Criterion A (impairment in personality functioning) and Criterion B (maladaptive traits) of the alternative DSM-5 model were assessed using self-report measures, dimensional and categorical PD classification were predicted from self-reported attachment style, mentalizing, childhood trauma, interparental conflict, and parental rejection. The ML algorithm correctly classified the presence of a PD in 91.01% of the cases. Sensitivity and specificity of ML predictions for categorical PD classification was 95.24% and 66.67%, respectively. ML predicted dimensional personality functioning on average within 0.67 standard deviations of the actual scores and effectively predicted levels of negative affect, detachment, and psychoticism. Attachment and mentalizing were the most important predictors of both Criterion A and B. The results indicate that ML can inform the assessment of PDs and advance research on personality theory.

Keywords: personality assessment, machine learning, personality disorder, personality functioning, maladaptive traits

Personality disorders (PDs) are very common in clinical and forensic samples (Beckwith, Moran, & Reilly, 2014; Fazel & Danesh, 2002) and represent one of the most prevalent mental disorders among the general population with a point prevalence of 12.16% (Volkert, Gablonski, & Rabung, 2018). Although PDs are associated with severe psychosocial impairments (American Psychiatric Association [APA], 2013; Huang et al., 2017; Narud, Mykletun, & Dahl, 2005), high economic burden (Soeteman, Hakkaart-van Roijen, Verheul, & Busschbach, 2008; Wagner et al., 2013), reduced life expectancy (Fok et al., 2012), and poorer treatment outcome of comorbid mental disorders (e.g. depression; Newton-Howes et al., 2014), PD diagnoses are rarely utilized in clinical practice (Oldham & Skodol, 1991; Zimmerman & Mattia, 1999) and PD assessment is regarded challenging (Biskin & Paris, 2012; Ekselius, 2018; Paris, 2007; Sarkar & Duggan, 2010; Tyrer, Reed, & Crawford, 2015).

The reasons for this are manifold; personality impairments interact with other mental disorders and hence are associated with high rates of comorbidity (McGlashan et al., 2000). Additionally, there are (so far) no biological markers that guide the PD diagnosis (Valencia Piedrahita & Cuartas Arias, 2016) and many patients see the cause of their problems in others instead of themselves (Klonsky, Oltmanns, & Turkheimer, 2002). Moreover, a large number of patients meet the criteria for more than one PD (Bernstein, 1998; Lilienfeld, Waldman, & Israel, 1994; McGlashan et al., 2000; Widiger & Trull, 1998), indicating that the current categorical PD diagnoses lack sufficient validity (Tyrer et al., 2015). In recent years, several surveys have shown that the majority of researchers and clinicians are therefore dissatisfied with the current categorical systems and advocate a dimensional model (Bernstein et al., 2007; Hansen et al., 2019; Hopwood et al., 2018; Morey, Skodol, & Oldham, 2014).

Dimensional PD Models

Emerging PD models such as the alternative model of the DSM-5 (AMPD; APA, 2013) and the rationale for ICD-11 (Tyrer et al., 2011; Tyrer et al., 2015) have gathered empirical support (Morey, Benson, Busch, & Skodol, 2015; Tyrer et al., 2014; Zimmermann, Kerber, Rek, Hopwood, & Krueger, 2019). Both models operationalize personality impairments on a single spectrum, resolving the issue of PD comorbidity (Bender, Morey, & Skodol, 2011; Tyrer, Mulder, Kim, & Crawford, 2019), allowing the assessment of impairment severity (including sub-threshold personality difficulties) and the presence of maladaptive personality traits (APA, 2013; Tyrer et al., 2019). Thus, both continuous models offer an evidence-based framework for research and clinical practice that address numerous shortcomings of categorical PD models (Tyrer et al., 2011; Zimmermann et al., 2019).

However, with the progressing implementation of dimensional models, several issues remain and continuous approaches pose new challenges for research on assessment and theory: Economic and reliable instruments to assess personality pathology are currently lacking (Tyer et al., 2015). Validated measures are usually very extensive and lengthy or depend on clinical experience (Tyrer et al., 2015), while shorter screening questionnaires tend to overdiagnose (Zimmerman, 1994). In addition, different methods and data sources (e.g. self & informant report, questionnaires & interviews) have shown only modest levels of agreement (Oltmanns & Oltmanns, 2019; Samuel, 2015). Moreover, researchers emphasize that current scientific theories of PDs are limited regarding scope, comprehensiveness or empirical support (Clarkin, 2018; Gunderson, Fruzzetti, Unruh, & Choi-Kain, 2018; Karterud & Kongerslev, 2019). Although various theories from different traditions have enhanced our comprehension of personality pathology and guided empirical research, the field is still lacking a unifying theory (Karterud & Kongerslev, 2019). Current dimensional models (such as the AMPD) are empirically derived and provide a framework for treating and studying PDs

but are atheoretical and thus not offer a conceptual integrative theory of PDs (Bach & Bernstein, 2019).

Therefore, new approaches and procedures are needed (1) to reliably assess PDs (Tyrer et al., 2015) and (2) to advance current theories and foster our understanding of personality pathology (Bleidorn & Hopwood, 2019).

Machine Learning

A promising approach that offers immense potential to clinical care and research is machine learning (ML; Bleidorn & Hopwood, 2019; Bzdok & Meyer-Lindenberg, 2018; Dwyer, Falkai, & Koutsouleris, 2019; Huys, Maia, & Frank, 2016; Iniesta, Stahl, & McGuffin, 2016). ML, or algorithmic methods, independently learns from data to find the best solution for complex problems (Dwyer et al., 2019). ML is a branch of artificial intelligence that describes computational strategies that automatically detect patterns from real data and generate solutions independently rather than being fixed a priori to a particular solution or underlying distribution (Dwyer et al., 2019). ML is a data driven procedure; it recognizes principles in sets of observations to make predictions, calculate their probability, and further optimizes the predictions by being able to build upon new, unseen data (Bzdok & Yeo, 2017). For this purpose, data sets are usually divided into subsamples; training samples are used to build a prediction model which is first evaluated in a validation sample utilizing techniques such as cross-validation, and then, finally, put to a rehearsal in a testing sample regarding its' predictive accuracy for completely new data (Dwyer et al., 2019; Iniesta et al., 2016).

ML (1) is theoretical agnostic (Huys et al., 2016), (2) allows quantitatively modeling of interactions between a nearly unlimited number of variables (Dwyer et al., 2019) and (3) different types of data (e.g. genes, physiology, behavior, and self-report; Dwyer et al., 2019) to identify relevant variables for a specific outcome. Moreover, ML techniques differs from

the “classical inferential paradigm” of modern psychological research that suffers from replication, reproducibility, p-value testing, procedural overfitting, and meaningless effect sizes (Dwyer et al., 2019). On the flip side, ML approaches are often criticized for being black-box-like, their predictions not being theory guided or interpretable as causal relationships (Castelvecchi, 2016).

Generalizability and accuracy of ML algorithms can be evaluated across different data sets or simulations that resample existing data (Iniesta et al., 2016) which allows better comparability of results. ML is not only able to model complex relationships and their impact on a specific variable but can inform multiple outcome variables (Bzdok & Meyer-Lindenberg, 2018). Applied to clinical care, such outcomes can be classification status (e.g. a diagnosis) (Arbabshirani, Plis, Sui, & Calhoun, 2017), drug dosage (Linden, Yarnold, & Nallamothu, 2016), treatment selection (Drysdale et al., 2017) or treatment prediction (Lee et al., 2018). Studies have shown that ML can make an important contribution to clinical practice and support clinical decision making: ML algorithms predict therapeutic outcomes of mood disorders (Lee et al., 2018), are able to diagnose various psychiatric disorders solely from brain data with high accuracy (Arbabshirani et al., 2017), and reliably diagnose cancer (e.g. Bejnordi et al., 2017). Regarding personality, recent research has applied ML to predict the “Big Five” personality traits from different materials, such as social media profiles and text messages (see Bleidorn & Hopwood, 2018 for an overview).

The Current Study

Researchers agree that ML has the potential to foster research and improve clinical care (Dwyer et al., 2019; Huys et al., 2016; Iniesta et al., 2016). Moreover, ML might be especially useful for cases where assessments are complex (Dwyer et al., 2018). However, no study has yet investigated the applicability of ML for PD assessment. This study has two aims: First, we

evaluate the usability of ML assessment of categorical and dimensional PD classification regarding specificity, sensitivity, and precision. Second, we investigate predictive patterns of PD classification in a continuous framework of personality pathology to identify empirically derived constituents of personality dysfunction. PDs are operationalized according to Criterion A (impairments in personality functioning) and Criterion B (maladaptive traits) of the AMPD (APA, 2013). PD predictions are based on constructs that are associated with PDs and/or are regarded as risk factors, namely insecure attachment, impaired mentalizing, childhood trauma, interparental conflict, and parental rejection.

Methods

Procedure

As studies on the developmental course of PDs have shown that symptoms peak in young age and decline in the course of adulthood (Alvarez-Tomás et al., 2017; Cohen, Crawford, Johnson, & Kasen, 2005; Zanarini, Frankenburg, Hennen, Reich, & Silk, 2005), we chose to recruit a sample of young adults. Inclusion criteria were age between 18 and 30 as well as sufficient knowledge of the German language. Participants were recruited via online platforms, flyers at a German University and social media to participate in a study on personality assessment. Data were collected online. After providing informed consent, participants were asked to complete a questionnaire battery, regarding personality functioning, maladaptive traits, mentalizing, attachment, childhood trauma, interparental conflict, and parental rejection. The respective measures are described below. The study was approved by the Ethics Commission of the Heidelberg-University (AZ Tau 2019 1/1).

Participants

670 began study participation; the study's final sample consisted of 410 young adults. Of those, 62.7% were female, 21.8% were male, and .6% reported no gender. Participant's age ranged between 18 and 30 with a mean age of 24.46 years (SD = 2.88). The sample showed a high level of education: 49.8% reported having an academic diploma, 48% a secondary school degree, and 2.2% no school degree at all. History of psychiatric illness was assessed through self-report: 13.5% had ever been affected by a mental disorder or had ever been in psychotherapeutic treatment. 11.6% were suffering from mental illness or were currently undergoing treatment at the time of participation in the study. The most frequent reported mental disorders were mood disorders (n = 33), trauma- and stressor-related disorders (n = 12), and PDs (n = 10).

Measures

Impairments in Personality Functioning. Only a few validated measures specifically designed for Criterion A are currently available (Hopwood, Good, & Morey, 2018). To assess personality impairments according to the alternative model, participants completed the Levels of Personality Functioning Scale – Self Report (LPFS-SR; Morey, 2017). This questionnaire consists of 80 items that are answered on a 4-point Likert scale (1 = Totally false, not at all true; 4 = Very true). Item scores are summed into four subscales, forming the four domains of Criterion A (identity, self-direction, empathy, and intimacy), and a total score (Morey, 2017). The measure has been evaluated in several samples (Hopwood, Good, & Morey, 2018) and is highly reliable (Hopwood, Good, & Morey, 2018; Morey, 2017). The German version (Müller & Zimmermann, 2018) is currently being validated (Zimmermann, personal communication, August 15, 2019). For this study, the dimensional total score and the categorical cut-off for a PD according to normative data of the LPFS-SR (Morey, 2017) were computed.

Maladaptive Traits. Criterion B of the AMPD was assessed with the Personality Inventory for DSM-5 Brief Form (PID-5-BF; APA, 2013), a self-report measure developed by the DSM-5 workgroup for PDs. The questionnaire consists of 25 items that are answered on a 4-point Likert scale (0 = Very false or often false; 3 = Very true or often true) and measure Criterion B's five maladaptive personality traits (negative affect, detachment, antagonism, disinhibition, and psychoticism) (APA, 2013). The German version has been validated in a clinical and nonclinical sample, demonstrating good psychometric properties (Zimmermann et al., 2014).

Attachment. Insecure attachment (anxiety and/or avoidance) is regarded as a key factor for the development of various mental disorders (Dozier, Stovall-McClough, & Albus, 2008) and a broad road range of studies have shown that attachment anxiety and/or attachment avoidance are associated with PDs (see Agrawal, Gunderson, Holmes, & Lyons-Ruth, 2004 for a review). We administered the German version of the Experiences in Close Relationships – Revised (ECR-RD; Ehrenthal, Dinger, Lamla, Funken, & Schauenburg, 2009) to assess attachment. The 36 items of the instrument are answered on a 6-point Likert scale (1 = Strongly disagree, 7 = Strongly agree) and load onto two subscales (attachment anxiety and avoidance). The German version displays good psychometric properties and has been validated in a clinical and nonclinical sample (Ehrenthal et al., 2009).

Mentalizing. Mentalizing is a crucial factor for understanding and treating PDs and a number of studies have shown that mentalizing is related to personality pathology (see Katznelson, 2014 for a review). Moreover, Criterion A and mentalization share a strong theoretical and empirical overlap (Bender et al., 2011; Zettl, Volkert, Vögele, Herpertz, Kubera, & Taubner, 2019). Mentalizing was assessed with the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016). The instrument assesses an individuals' certainty and uncertainty about mental states, the 8 items are answered on a 7-point Likert scale (0 =

Strongly disagree, 5 = Strongly agree), and are organized into two subscales (RFQ Certainty, RFQ Uncertainty) (Fonagy et al., 2016). The measure was validated in several clinical and non-clinical samples, demonstrating sufficient internal consistency and test-retest-reliability (Fonagy et al., 2016). The German version was retrieved from the authors but has yet to undergo validation.

Childhood Trauma. Adverse childhood experiences increase the risk for developing a PD and longitudinal studies link childhood trauma to personality pathology (Björkenstam, Ekselius, Burstrom, Kosidou, & Björkenstam, 2017). In this study, we assessed childhood trauma with the German version of the Childhood Trauma Questionnaire (Wingenfeld et al., 2010). The questionnaire features 28 items, answered on a 5-point Likert scale (1 = Never true; Very often true), that assess five types of trauma (emotional abuse and neglect, physical abuse and neglect, and sexual abuse) (Wingenfeld et al., 2010). The German version was validated in a clinical and representative sample, showing good factorial and convergent validity and high internal consistency (Wingenfeld et al., 2010; Klinitzke, Romppel, Häuser, Brähler, & Glaesmer, 2011).

Interparental Conflict. Family and interparental conflict, maternal-child discord, and parent-child relationship have been shown to be associated with PDs and psychosocial functioning (Bezirganian, Cohen, & Brook, 1993; Boucher et al., 2017; Stepp, Olino, Klein, Seeley, & Lewinsohn, 2013). To assess parental conflict, we administered the German short version of the Children's Perception of Interparental Conflict Scale (Gödde & Walper, 2001). The measure comprises 15 items that are answered on a 5-point Likert scale (1 = Never; 5 = Very often) (Gödde & Walper, 2001). The subscales assess five aspects of parental conflicts (frequency, harmony, unharmony, child as mediator of conflict, and child as origin of conflict) (Gödde & Walper, 2001). The German Version has been validated in a sample of

children and youths, demonstrating good internal consistency and good validity (Gödde & Walper, 2001).

Parental Rejection. Negative parenting practices such as low warmth and perceived parental rejection are associated with increased PD symptoms and personality maladjustment (Reinelt et al., 2014; Khaleque, 2017; Stepp, Lazaus, & Byrd, 2016). In this study we assessed Parental acceptance and rejection with the short version of the Parental Acceptance-Rejection Questionnaire (Rohner & Khaleque, 2005). The questionnaire measures retrospective memories of rejection and acceptance by parents in childhood. The short version consists of 24 items that are answered on a 4-point Likert scale (1 = Never applies; 4 = Almost always applies). The questionnaire is available in 52 languages and has been proven to be reliable and valid in over 51 studies but has yet not been validated in a German population (Khaleque & Rohner, 2002).

Statistical Analyses

Since the study data was gathered via an online questionnaire employing forced choice questions for the most part, the data contained merely 1.39% missing values. Imputation was done assuming an observation's missingness not to be related to the dependent variable at dropout (Enders, 2011). Multiple imputations by chained equations (MICE, van Buuren & Groothuis-Oudshoorn 2008), using fully conditional specification with 10 iterations was used to impute missing values. MICE produces asymptotically unbiased estimations of the data under these missingness assumptions (White, Royston, & Wood 2011).

To investigate, which study variable best predicted personality functioning according to the LPFS-SR, categorical PD classification according to the LPFS-SR, as well as maladaptive personality traits according to the PID-5-BF (negative affect, detachment, antagonism, disinhibition, and psychoticism), Gradient Boosting Machines (GBM) using

regression trees was employed. GBM are a form of ensemble learning. The base learners in a GBM are so called “weak learners” that are trained sequentially. GBM starts with an initial model for the data, in our case a single regression tree, and constructs a new model by successively fitting the residuals of the current model rather than the outcome (James, 2013). It learns slowly by specifically targeting the areas of the data where the prior models do not do well. GBM, specifically XGBoost, was utilized because it is insensitive to outliers and assumes no distribution in the outcome or underlying data mechanism while being the ensemble learning method of choice (Breiman, 2001; Hastie, 2009; Chen, 2016).

The study data was split on the respective outcome 70% - 30% into a training and test set. The algorithms were trained on the training set using 5 fold cross-validation with 10 repeats in a grid search for the optimal hyperparameters. The hyperparameter grid search for the GBM was done by iteratively manipulating the shrinkage coefficient η between 0.01 and 0.2, the interaction depth of each tree (max_depth) between 1 and 6, the number of boosting iterations (nrounds) between 1 and 1000, while keeping the minimum loss reduction (gamma) fixed at 0 and the minimum sum of instance weight (min_child_weight) fixed at 1. This represents a conservative approach with little likelihood of overfitting. The algorithms with the highest prediction accuracy in the grid search were then chosen for the final validation in the test set. We used the root mean squared error (RMSE) as accuracy metric for the continuous outcomes and prediction accuracy as metric for. Marginal effects were calculated utilizing the tree traversal method developed by Friedman (2001). Relative variable importance for the models was computed as the relative influence of the variable on the reduction in the loss function of the GBM.

All statistical analyses were performed using R version 3.6.0 (R Core Team, 2008). The R package “caret” version 6.0-84 (Kuhn, 2008) was used to train the algorithms.

Results

Descriptive statistics and Cronbach's alpha values for all study variables are listed in Table 1. According to the participants' LPFS-SR scores, dimensional PD status was as follows: 74.88% = little or no impairment, 10.49% = personality difficulty, 14.63% = personality impairment.

Personality Functioning

Categorical PD Status. The values for the hyperparameters of the final model predicting categorical PD status according to the LPFS-SR were 200 boosting iterations at a depth of 1 with $\eta = 0.1$. A training set accuracy of 95.43% was achieved. The test set accuracy was 91.06%, at a no information rate (NIR) of 85.37% (p accuracy > NIR = 0.042) with a sensitivity of 95.24% and a specificity of 66.67%. Figure 1a shows the marginal effects of the two most important variables; Figure 1b the most important variables of the model. Both ECR Anxiety and ECR Avoidance showed an extreme plateau effect. Values of below 60 for ECR Avoidance and below 70 for ECR Anxiety had no effect on the probability of being classified with a PD. Above those values however, the probability of being classified with a PD increases fast with the highest probabilities being observed at levels of avoidance and anxiety above 70.

Dimensional Personality Functioning Score. The values for the hyperparameters of the final model predicting LPFS-SR total were 200 boosting iterations at a depth of 1 with $\eta = 0.1$. This resulted in a within training set accuracy of RMSE = 42.55 with an R^2 of 0.60. The test set accuracy was RMSE = 46.10 with an R^2 of 0.57. This means the models predictions were on average within 0.67 standard deviations of the actual scores. Figure 1c shows the marginal effects of the two most important variables; Figure 1d the most important variables of the model. Both ECR Anxiety and RFQ Certainty showed a relative linear association with

the predicted LPFS-SR value, although both variables had little effect in the first halves of their range.

Maladaptive Traits

Negative Affect. The values for the hyperparameters of the final model predicting the PID-5-BF subscale Negative Affect were 60 boosting iterations at a depth of 1 with $\eta = 0.2$. This resulted in a within training set accuracy of RMSE = 2.38 with an R^2 of 0.44. The test set accuracy was RMSE = 2.41 with an R^2 of 0.41. This means the models predictions were on average within 0.75 standard deviations of the actual scores. Figure 2a shows the marginal effects of the two most important variables; Figure 2b the most important variables of the model. Both RFQ Uncertainty and ECR Anxiety show a nearly linear relationship with the predicted outcome.

Detachment. The values for the hyperparameters of the final model predicting the PID-5-BF subscale Detachment were 100 boosting iterations at a depth of 1 with $\eta = 0.1$. This resulted in a within training set accuracy of RMSE = 2.27 with an R^2 of 0.43. The test set accuracy was RMSE = 2.38 with an R^2 of 0.41. This means the models predictions were on average within 0.79 standard deviations of the actual scores. Figure 2c shows the marginal effects of the two most important variables; Figure 2d the most important variables of the model. Both RFQ Uncertainty and ECR Avoidance show a nearly linear relationship with the predicted outcome.

Psychoticism. The values for the hyperparameters of the final model predicting the PID-5-BF subscale Psychoticism were 60 boosting iterations at a depth of 1 with $\eta = 0.1$. This resulted in a within training set accuracy of RMSE = 2.23 with an R^2 of 0.42. The test set accuracy was RMSE = 2.42 with an R^2 of 0.38. This means the models predictions were on average within 0.80 standard deviations of the actual scores. Figure 2e shows the marginal

effects of the two most important variables; Figure 2f the most important variables of the model. RFQ Certainty displays a short plateau at with values of above 2.25 having no effect on the predicted value. For RFQ Certainty values below 2.25 the association with the predicted outcome is nearly linear. ECR Avoidance on the other hand shows a concise threshold at an ECR Avoidance value of 75 with values above that being associated with a steep increase in predicted psychoticism.

Antagonism & PID Disinhibition. For the PID-5-BF subscales Antagonism and Disinhibition the hyperparameter grid search did not find a model that predicted the training set observation with a RMSE < 1 standard deviation of the outcome. We thus judged the resulting estimates from such a model to be too vague and did not choose a final model for the two subscales.

Discussion

In this study, we (1) tested ML as a method for PD assessment with regard to personality functioning and maladaptive traits and (2) investigated ML-derived predictors of personality pathology. We assessed Criterion A and B of the AMPD with validated self-report questionnaires and conducted ML to predict personality functioning, maladaptive traits, as well as categorical PD classification. Predictions were based on a number of clinically relevant factors that are associated and/or are regarded as a risk factor for the development of PDs (namely insecure attachment, impaired mentalizing, childhood trauma, interparental conflict, and parental rejection). In the following we discuss our main findings and discuss limitations as well as future directions for research.

ML Predictions of Personality Functioning

The ML algorithm applied in this study demonstrated high precision in the prediction of continuous personality functioning. Predictions were on average within 0.67 standard deviations of the participants' actual scores. This is a promising accuracy, considering that personality functioning was predicted largely on the basis of attachment anxiety and mentalizing. Because this is the first study to apply ML techniques to PD assessment, the observed precision cannot be compared yet with previous research.

Regarding the presence of a PD, the algorithm correctly classified about 91% of the participants. The probability that a participant with a PD was correctly identified by the ML algorithm was 95%. Participants not affected with a PD were correctly identified with a 66% probability. Consequently, there is a five-percent probability not to receive a diagnosis although a PD is present, and a nearly 44-percent probability of receiving a PD diagnosis without having a disorder. The ML algorithm was more sensitive than specific and thus might be useful for detecting individuals with a PD, but not that useful for detecting non-cases. The especially strong sensitivity potentially qualifies ML as an alternative screening method for detecting (clinical as well as sub-clinical) personality impairments; sensitivity and specificity of ML-predicted categorical PD status are comparable with screening questionnaires for PDs: A meta-analytic review by Gárriz & Gutiérrez (2009) showed that sensitivity and specificity of various PD measures was on average .80 and .73, respectively.

ML Predictions of Maladaptive Traits

The ML algorithm predicted three out of the 5 maladaptive traits with sufficient precision. Negative affect, detachment, and psychoticism were predicted with accuracy similar as for personality functioning, whereas the algorithm did not find an accurate prediction model for antagonism and disinhibition. However, as for personality functioning, the precision of ML predictions cannot be compared with previous studies as there are no further studies yet.

Regarding disinhibition and antagonism, the algorithm failed to build a prediction model with sufficient precision. This demonstrates why ML is often referred to as a black box, because the algorithm is not able to explain why no solution was found. Two explanations seem plausible: as attachment and mentalizing were the most predominant predictors across the other facets, it might be that disinhibition and antagonism can best be explained by other variables that were not included in this study. Second, the sample showed a significantly lower variance regarding these two facets. Thus, the algorithm may not have been able to determine an accurate prediction model.

Predictors of Personality Pathology

Attachment and *mentalizing* were consistent predictors elicited by the ML algorithm, not solely of personality functioning but also of each of the 3 maladaptive traits. The pattern of results is in line with a novel theory of PDs, the Temperament-Attachment-Mentalization Theory (TAM; Karterud & Kongerslev, 2019), as well as with mentalization-centered theories (Fonagy, Luyten, & Allison, 2015; Fonagy, Luyten & Strathearn, 2011). In comparison to attachment and mentalizing, childhood trauma, interparental conflicts (related specifically to the child's perception of being the origin of parental conflicts) as well as maternal rejection significantly added to the prediction of personality functioning (i.e. Criterion A), but to a far lesser degree. Most studies on childhood neglect and abuse support a rather moderate and heterogenous association between personality pathology and adverse childhood experiences (Fossati, Madeddu, & Maffei, 1999).

However, for negative affect, detachment, and psychoticism (i.e. Criterion B), the patterns differ for each of the facets: *Negative affect* was mainly predicted by attachment anxiety and uncertainty about mental states, whereas attachment avoidance and certainty were key predictors of *detachment*. Levels of *psychoticism*, on the other hand, were best predicted

by certainty about mental states and attachment anxiety. In detail, different components of the two constructs are decisive for the prediction of maladaptive traits. As attachment anxiety promotes emotional hypersensitivity and reactiveness to social stimuli (Kobak & Sceery, 1988), frequent and intense negative emotions such as anxiety, as defined by negative affect (Krueger & Markon, 2014), may be more strongly pronounced in such individuals. Consequently, attachment avoidance, enhancing distance of oneself from others and their emotions (Kobak & Sceery, 1988), promotes avoidance of socioemotional experiences, as defined by detachment (Krueger & Markon, 2014).

Our current results therefore not only yield support for the utility of ML for PD assessment, but we also replicate relevant empirical support for the theoretical overlap between attachment, mentalization and personality pathology as operationalized in the AMPD (Bender et al., 2013; Karterud & Kongersley, 2019; Zettl et al., 2019).

Progressing in the assessment of PDs

Implementing ML for PD diagnostics could yield many advantages: (1) Assessments can be based on just a few established instruments, as ML can even be applied on an item-level. Consequently, PD assessments would be less time consuming. (2) Results of ML-predictions are easy to interpret as decisive predictors, accuracy and probability are determined by the algorithm. (3) Empirical research on risk factors and the pathogenesis of PDs could be directly translated into clinical care by integrating corresponding measures into the prediction model. (4) The benefits of ML are not limited to PD assessment alone but, with further research, could also inform treatment planning or treatment selection and predict the course of treatment.

Although the generalizability of our results is limited, the study provides first evidence that ML can be applied to PD assessment. The algorithm was able to achieve a correct PD

classification in over 90% of the cases and sensitivity and specificity are approximately on a par with PD screening questionnaires. This is a promising accuracy, given of the predictors in this study. However, to make ML useful for clinical care beyond existing screening questionnaires, further progress is needed. Moving forward, ML algorithms could be applied to items of different PD questionnaires (self- and informant-report) as well as expert ratings of PD severity to form a clinical support system that not only yields high accuracy but also high validity. However, to gauge the full potential of ML for PD assessment must be the subject of future studies.

Limitations and Future Directions

The results of our study must be considered in the light of several limitations. First, the sample demonstrated a predominantly high level of personality functioning and low variance of disinhibition and antagonism. Although the entire spectrum of Criterion A was covered, higher severities of PD impairment were clearly underrepresented. As a result, predicting more severe levels of personality functioning is more difficult for the ML algorithm, because less cases for training and validation is available. Furthermore, probably due to the lack of variance regarding two of the maladaptive facets, the algorithm was not able to build a precise prediction model for disinhibition and antagonism. In addition, the absence of a dedicated clinical sample limits the generalizability of the results. Future studies are needed to evaluate the usability of ML in clinical samples with validated PD diagnoses. Given that ML performs better the more data available, ML predictions can achieve even better results with sufficient data from psychiatric samples.

Second, the study had only one outcome measure for personality functioning and one for maladaptive traits, which we considered as the gold standard/true scores for all analyses. Although both questionnaires have been validated in several samples and directly correspond

to Criterion A and B, our interpretations are limited to the applied measures. Therefore, we recommend incorporating several outcome measures for future studies. This yields two advantages: First, the validity and generalizability of ML predictions can be tested across multiple measures for PDs. Second, PDs can in turn be predicted from a number of measures or items for PD assessment, which should allow much greater accuracy in future studies.

Third, PD classifications were predicted only from a small amount of information. We did not assess socioeconomic status, Axis I disorders or emotion dysregulation, a core concept of borderline PD (Carpenter & Trull, 2013; Selby & Joiner, 2009). As comorbidity is the rule rather than the exception, future studies should systematically screen for comorbid mental disorders to facilitate comparability of results across several samples. In addition, data from clinical trials and cohort studies are needed to advance ML for PD assessment, as demographic, clinical and health record information could be used to further improve PD predictions.

Fourth, we recommend for future studies to adopt a multi-method approach by gathering data from different formats and sources. Our analyses are based exclusively on self-report which limit the generalizability, as meta-analyses have shown that self-other agreement of PD traits and symptoms is low to moderate. Therefore, further research is needed to apply ML to self-, informant-, and/or clinician-reports of PD pathology. The fact that ML is able to model complex patterns between unlimited quantities of variables to predict multiple outcomes at once opens up the potential to further investigate differences in self-other agreement of PD assessment.

Conclusion

ML provides a framework for solving complex problems that can be applied to inform PD assessment and advance personality theory. In this study, we were able to predict level of

personality functioning, categorical PD classification, and three maladaptive traits from various indicators of psychopathology. Insecure attachment and impaired mentalizing were crucial predictors of personality pathology. The results of this study provide first evidence that ML can be used to assist PD assessment. Moreover, we add empirical evidence to the novel TAM theory of personality pathology. However, further research is needed to evaluate the generalizability of our results and to gauge the full potential of ML.

References

Agrawal, H. R., Gunderson, J., Holmes, B. M., & Lyons-Ruth, K. (2004). Attachment studies with borderline patients: A review. *Harvard Review of Psychiatry*, *12*(2), 94-104. https://doi.org/10.1080/10673220490447218

Alvarez-Tomás, I., Soler, J., Bados, A., Martín-Blanco, A., Elices, M., Carmona, C., ... & Pascual, J. C. (2017). Long-term course of borderline personality disorder: a prospective 10-year follow-up study. *Journal of Personality Disorders*, *31*(5), 590-605. https://doi.org/10.1521/pedi_2016_30_269

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: Author.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage*, *145*, 137-165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Bach, B., & Bernstein, D. P. (2019). Schema therapy conceptualization of personality functioning and traits in ICD-11 and DSM-5. *Current Opinion in Psychiatry*, *32*(1), 38-49. https://doi.org/10.1097/YCO.0000000000000464

Beckwith, H., Moran, P. F., & Reilly, J. (2014). Personality disorder prevalence in psychiatric outpatients: a systematic literature review. *Personality and Mental Health*, *8*(2), 91-101. https://doi.org/10.1002/pmh.1252

Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... & Geessink, O. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, *318*(22), 2199-2210. https://doi.org/10.1001/jama.2017.14585

Bender, D. S., Morey, L. C., & Skodol, A. E. (2011). Toward a model for assessing level of personality functioning in DSM–5, part I: A review of theory and methods. *Journal of Personality Assessment*, *93*(4), 332-346. https://doi.org/10.1080/00223891.2011.583808

Bernstein, R. F. (1998). Reconceptualizing personality disorder diagnosis in the DSM-V: the discriminant validity challenge. *Clinical Psychology: Science and Practice*, *5*(3), 333-343. https://doi.org/10.1111/j.1468-2850.1998.tb00153.x

Bernstein, D. P., Iscan, C., Maser, J., & Boards of Directors of the Association for Research in Personality Disorders and the International Society for the Study of Personality Disorders. (2007). Opinions of personality disorder experts regarding the DSM-IV personality disorders classification system. *Journal of Personality Disorders*, *21*(5), 536-551. https://doi.org/10.1521/pedi.2007.21.5.536

Bezirganian, S., Cohen, P., & Brook, J. S. (1993). The impact of mother-child interaction on the development of borderline personality disorder. *The American Journal of Psychiatry, 150*(12), 1836-1842.https://doi.org/10.1176/ajp.150.12.1836

Biskin, R. S., & Paris, J. (2012). Diagnosing borderline personality disorder. *CMAJ*, *184*(16), 1789-1794. https://doi.org/10.1503/cmaj.090618

Björkenstam, E., Ekselius, L., Burström, B., Kosidou, K., & Björkenstam, C. (2017). Association between childhood adversity and a diagnosis of personality disorder in young adulthood: a cohort study of 107,287 individuals in Stockholm County. *European Journal of Epidemiology*, *32*(8), 721-731. https://doi.org/10.1007/s10654-017-0264-9

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*(2), 190-203. https://doi.org/10.1177/1088868318772990

Boucher, M. È., Pugliese, J., Allard-Chapais, C., Lecours, S., Ahoundova, L., Chouinard, R., & Gaham, S. (2017). Parent–child relationship associated with the development of borderline personality disorder: a systematic review. *Personality and Mental Health*, *11*(4), 229-255. https://doi.org/10.1002/pmh.1385

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bzdok, D., & Yeo, B. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *Neuroimage*, *155*, 549-564. https://doi.org/10.1016/j.neuroimage.2017.04.061

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223-230. https://doi.org/10.1016/j.bpsc.2017.11.007

Carpenter, R. W., & Trull, T. J. (2013). Components of emotion dysregulation in borderline personality disorder: A review. *Current Psychiatry Reports*, *15*(1), 335. https://doi.org/10.1007/s11920-012-0335-2

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News*, *538*(7623), 20-23.

Chen, T., & Guestrin, C. (2016). XGBoost. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16* (pp. 785–794). New York, New York, USA: ACM Press. https://doi.org/10.1145/2939672.2939785

Clarkin, J. F. (2018). Theories and Treatment of Borderline Personality Disorder: Commentary on Gunderson et al. *Journal of Personality Disorders*, *32*(2), 175-180. https://doi.org/10.1521/pedi.2018.32.2.175

Cohen, P., Crawford, T. N., Johnson, J. G., & Kasen, S. (2005). The children in the community study of developmental course of personality disorder. *Journal of Personality Disorders*, *19*(5), 466-486. https://doi.org/10.1521/pedi.2005.19.5.466

Dozier, M., Stovall-McClough, K. C., & Albus, K. E. (2008). Attachment and psychopathology in adulthood. In J. Cassidy & P. R. Shaver (Eds.), *Handbook of attachment: Theory, research, and clinical applications* (pp. 718-744). New York, NY, US: The Guilford Press.

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., ... & Schatzberg, A. F. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, *23*(1), 28-38. https://doi.org/10.1038/nm.4246

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91-118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Ehrenthal, J. C., Dinger, U., Lamla, A., Funken, B., & Schauenburg, H. (2009). Evaluation of the German Version of the Attachment Questionnaire „Experiences in Close Relationships – Revised" (ECR-RD). *PPmP-Psychotherapie· Psychosomatik· Medizinische Psychologie*, *59*(06), 215-223. https://doi.org/10.1055/s-2008-1067425

Ekselius, L. (2018). Personality disorder: a disease in disguise. *Upsala Journal of Medical Sciences*, *123*(4), 194-204. https://doi.org/10.1080/03009734.2018.1526235

Enders, C. K. (2010). *Applied missing data analysis*. *Methodology in the social sciences*. New York NY u.a.: Guilford.

Fazel, S., & Danesh, J. (2002). Serious mental disorder in 23 000 prisoners: a systematic review of 62 surveys. *The Lancet*, *359*(9306), 545-550. https://doi.org/10.1016/S0140-6736(02)07740-1

Fok, M. L. Y., Hayes, R. D., Chang, C. K., Stewart, R., Callard, F. J., & Moran, P. (2012). Life expectancy at birth and all-cause mortality among people with personality disorder. *Journal of Psychosomatic Research*, *73*(2), 104-107. https://doi.org/10.1016/j.jpsychores.2012.05.001

Fonagy, P., Luyten, P., & Strathearn, L. (2011). Borderline personality disorder, mentalization, and the neurobiology of attachment. *Infant Mental Health Journal*, *32*(1), 47-69. https://doi.org/10.1002/imhj.20283

Fonagy, P., Luyten, P., & Allison, E. (2015). Epistemic petrification and the restoration of epistemic trust: A new conceptualization of borderline personality disorder and its

psychosocial treatment. *Journal of Personality Disorders*, *29*(5), 575-609. https://doi.org/10.1521/pedi.2015.29.5.575

Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y. W., Warren, F., Howard, S., ... & Lowyck, B. (2016). Development and validation of a self-report measure of mentalizing: The reflective functioning questionnaire. *PLoS One*, *11*(7), e0158678. https://doi.org/10.1371/journal.pone.0158678

Fossati, A., Madeddu, F., & Maffei, C. (1999). Borderline personality disorder and childhood sexual abuse: a meta-analytic study. *Journal of Personality Disorders*, *13*(3), 268-280. https://doi.org/10.1521/pedi.1999.13.3.268

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gárriz, M., & Gutiérrez, F. (2009). Personality disorder screenings: A meta-analysis. *Actas Esp Psiquiatr*, *37*(3), 148-52.

Gunderson, J. G., Fruzzetti, A., Unruh, B., & Choi-Kain, L. (2018). Competing theories of borderline personality disorder. *Journal of Personality Disorders*, *32*(2), 148-167. https://doi.org/10.1521/pedi.2018.32.2.148

Gödde, M., & Walper, S. (2001). The German short version of the Children's Perception of Interparental Conflict Scale. *Diagnostica, 47*(1), 18-26. https://doi.org/10.1026//0012-1924.47.1.18

Hansen, S. J., Christensen, S., Kongerslev, M. T., First, M. B., Widiger, T. A., Simonsen, E., & Bach, B. (2019). Mental health professionals' perceived clinical utility of the ICD-10 vs. ICD-11 classification of personality disorders. *Personality and Mental Health*, *13*(2), 84-95. https//doi.org/10.1002/pmh.1442

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hopwood, C. J. (2018). A framework for treating DSM-5 alternative model for personality disorder features. *Personality and Mental Health*, *12*(2), 107-125. https://doi.org/10.1002/pmh.1414

Hopwood, C. J., Good, E. W., & Morey, L. C. (2018). Validity of the DSM–5 levels of personality functioning scale–self report. *Journal of Personality Assessment*, *100*(6), 650-659. https://doi.org/10.1080/00223891.2017.1420660

Hopwood, C. J., Kotov, R., Krueger, R. F., Watson, D., Widiger, T. A., Althoff, R. R., ... & Bornovalova, M. A. (2018). The time has come for dimensional personality disorder diagnosis. *Personality and Mental Health*, *12*(1), 82-86. https://doi.org/10.1002/pmh.1408

Huang, I. C., Lee, J. L., Ketheeswaran, P., Jones, C. M., Revicki, D. A., & Wu, A. W. (2017). Does personality affect health-related quality of life? A systematic review. *PloS One*, *12*(3), e0173806. https://doi.org/10.1371/journal.pone.0173806

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404-413. https://doi.org/10.1038/nn.4238

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455-2465. https://doi.org/10.1017/S0033291716001367

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Karterud, S. W., & Kongerslev, M. T. (2019). A temperament-attachment-mentalization-based (TAM) theory of personality and its disorders. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00518

Katznelson, H. (2014). Reflective functioning: A review. *Clinical Psychology Review*, *34*(2), 107-117. https://doi.org/10.1016/j.cpr.2013.12.003

Khaleque, A., & Rohner, R. P. (2002). Reliability of measures assessing the pancultural association between perceived parental acceptance-rejection and psychological adjustment: A meta-analysis of cross-cultural and intracultural studies. *Journal of Cross-Cultural Psychology*, *33*(1), 87-99. https://doi.org/10.1177/0022022102033001006

Khaleque, A. (2017). Perceived parental hostility and aggression, and children's psychological maladjustment, and negative personality dispositions: a meta-analysis. *Journal of Child and Family Studies*, *26*(4), 977-988. https://doi.org/10.1007/s10826-016-0637-9

Klinitzke, G., Romppel, M., Häuser, W., Brähler, E., & Glaesmer, H. (2012). The German Version of the Childhood Trauma Questionnaire (CTQ) – Psychometric Characteristics in a Representative Sample of the General Population. *PPmP-Psychotherapie· Psychosomatik· Medizinische Psychologie*, *62*(02), 47-51. https://doi.org/10.1055/s-0031-1295495

Klonsky, E. D., Oltmanns, T. F., & Turkheimer, E. (2002). Informant-reports of personality disorder: Relation to self-reports and future research directions. *Clinical Psychology: Science and Practice, 9*(3), 300-311. http://doi.org/10.1093/clipsy/9.3.300

Kobak, R. R., & Sceery, A. (1988). Attachment in late adolescence: Working models, affect regulation, and representations of self and others. *Child Development*, *59*(1) 135-146. https://doi.org/10.2307/1130395

Krueger, R. F., & Markon, K. E. (2014). The role of the DSM-5 personality trait model in moving toward a quantitative and empirically based approach to classifying personality and psychopathology. *Annual Review of Clinical Psychology*, *10*, 477-501. https://doi.org/10.1146/annurev-clinpsy-032813-153732

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

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., ... & Chan, T. C. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519-532. https://doi.org/10.1016/j.jad.2018.08.073

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice*, *1*(1), 71-83. https://doi.org/10.1111/j.1468-2850.1994.tb00007.x

Linden, A., Yarnold, P. R., & Nallamothu, B. K. (2016). Using machine learning to model dose–response relationships. *Journal of Evaluation in Clinical Practice*, *22*(6), 860-867. https://doi.org/10.1111/jep.12573

McGlashan, T. H., Grilo, C. M., Skodol, A. E., Gunderson, J. G., Shea, M. T., Morey, L. C., ... & Stout, R. L. (2000). The collaborative longitudinal personality disorders study: Baseline axis I/II and II/II diagnostic co-occurrence. *Acta Psychiatrica Scandinavica*, *102*(4), 256-264. https://doi.org/ 10.1034/j.1600-0447.2000.102004256.x

Morey, L. C., Skodol, A. E., & Oldham, J. M. (2014). Clinician judgments of clinical utility: A comparison of DSM-IV-TR personality disorders and the alternative model for DSM-5 personality disorders. *Journal of Abnormal Psychology*, *123*(2), 398. http://doi.org/10.1037/a0036481

Morey, L. C., Benson, K. T., Busch, A. J., & Skodol, A. E. (2015). Personality disorders in DSM-5: Emerging research on the alternative model. *Current Psychiatry Reports*, *17*(4), 24. https://doi.org/10.1007/s11920-015-0558-0

Morey, L. C. (2017). Development and initial evaluation of a self-report form of the DSM–5 Level of Personality Functioning Scale. *Psychological Pssessment*, *29*(10), 1302-1308. https://doi.org/10.1037/pas0000450

Müller, S. & Zimmermann, J. (2018). Skala zur Erfassung des Funktionsniveaus der Persönlichkeit – Selbsteinschätzung (SEFP-S). Deutsche Übersetzung des LPFS-SR. Unpublished manuscript.

Narud, K., Mykletun, A., & Dahl, A. A. (2005). Quality of life in patients with personality disorders seen at an ordinary psychiatric outpatient clinic. *BMC Psychiatry*, *5*(1), 10. https://doi.org/10.1186/1471-244X-5-10

Newton-Howes, G., Tyrer, P., Johnson, T., Mulder, R., Kool, S., Dekker, J., & Schoevers, R. (2014). Influence of personality on the outcome of treatment in depression: systematic

review and meta-analysis. *Journal of Personality Disorders*, *28*(4), 577-593. https://doi.org/10.1521/pedi_2013_27_070

Oldham, J. M., & Skodol, A. E. (1991). Personality disorders in the public sector. *Psychiatric Services*, *42*(5), 481-487. https://doi.org/10.1176/ps.42.5.481

Oltmanns, J. R., & Oltmanns, T. F. (2019). Self–other agreement on ratings of personality disorder symptoms and traits: Three meta-analyses. In T. D. Letzring and J. S. Spain (Eds.), *The Handbook of Accurate Personality Judgment: Theory and Empirical Findings*. Oxford University Press.

Paris, J. (2007). Why psychiatrists are reluctant to diagnose: borderline personality disorder. *Psychiatry (Edgmont)*, *4*(1), 35-39.

Preti, E., Di Pierro, R., Costantini, G., Benzi, I. M., De Panfilis, C., & Madeddu, F. (2018). Using the Structured Interview of Personality Organization for DSM–5 Level of Personality Functioning Rating Performed by Inexperienced Raters. *Journal of Personality Assessment*, *100*(6), 621-629. https://doi.org/10.1080/00223891.2018.1448985

Reinelt, E., Stopsack, M., Aldinger, M., Ulrich, I., Grabe, H. J., & Barnow, S. (2014). Longitudinal transmission pathways of borderline personality disorder symptoms: from mother to child?. *Psychopathology*, *47*(1), 10-16. https://doi.org/10.1159/000345857

Rohner, R. P., & Khaleque, A. (2005). Parental acceptance-rejection questionnaire (PARQ): Test manual. *Handbook for the Study of Parental Acceptance and Rejection*, *4*, 43-106.

R Development Core Team. (2008). *R: A language and environment for statistical computing*: Vienna. Retrieved from http://www.R-project.org

Samuel, D. B. (2015). A review of the agreement between clinicians' personality disorder diagnoses and those from other methods and sources. *Clinical Psychology: Science and Practice*, *22*(1), 1-19. https://doi.org/10.1111/cpsp.12088

Sarkar, J., & Duggan, C. (2010). Diagnosis and classification of personality disorder: difficulties, their resolution and implications for practice. *Advances in Psychiatric Treatment*, *16*(5), 388-396. https://doi.org/10.1192/apt.bp.108.006015

Selby, E. A., & Joiner Jr, T. E. (2009). Cascades of emotion: The emergence of borderline personality disorder from emotional and behavioral dysregulation. *Review of General Psychology*, *13*(3), 219-229. https://doi.org/10.1037/a0015687

Stepp, S. D., Olino, T. M., Klein, D. N., Seeley, J. R., & Lewinsohn, P. M. (2013). Unique influences of adolescent antecedents on adult borderline personality disorder features. *Personality Disorders: Theory, Research, and Treatment, 4*(3), 223-229. https://doi.org/10.1037/per0000015

Stepp, S. D., Lazarus, S. A., & Byrd, A. L. (2016). A systematic review of risk factors prospectively associated with borderline personality disorder: Taking stock and moving forward. *Personality Disorders: Theory, Research, and Treatment, 7*(4), 316-323. https://doi.org/10.1037/per0000186

Soeteman, D. I., Roijen, L. H. V., Verheul, R., & Busschbach, J. J. (2008). The economic burden of personality disorders in mental health care. *Journal of Clinical Psychiatry*, *69*(2), 259-265. https://doi.org/10.4088/JCP.v69n0212

Tyrer, P., Crawford, M., Mulder, R., Blashfield, R., Farnam, A., Fossati, A., ... & Swales, M. (2011). The rationale for the reclassification of personality disorder in the 11th revision of the International Classification of Diseases (ICD-11). *Personality and Mental Health*, *5*(4), 246-259. https://doi.org/10.1002/pmh.190

Tyrer, P., Crawford, M., Mulder, R., Blashfield, R., Farnam, A., Fossati, A., & Reed, G. M. (2011). A classification based on evidence is the first step to clinical utility. *Personality and Mental Health*, *5*(4), 304-307. http://doi.org/10.1002/pmh.189

Tyrer, P., Crawford, M., Sanatinia, R., Tyrer, H., Cooper, S., Muller-Pollard, C., ... & Guo, B. (2014). Preliminary studies of the ICD-11 classification of personality disorder in practice. *Personality and Mental Health*, *8*(4), 254-263. https://doi.org/10.1002/pmh.1275

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet*, *385*(9969), 717-726. https://doi.org/10.1016/S0140-6736(14)61995-4

Tyrer, P., Mulder, R., Kim, Y. R., & Crawford, M. J. (2019). The development of the ICD-11 classification of personality disorders: an amalgam of science, pragmatism, and politics. *Annual Review of Clinical Psychology*, *15*, 481-502. https://doi.org/10.1146/annurev-clinpsy-050718-095736

Valencia Piedrahita, M., & Cuartas Arias, J. (2016). Potential biomarkers in personality disorders: current state and future research. *International Journal of Psychological Research*, *9*(1), 98-112. https://doi.org/10.21500/20112084.2105

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Volkert, J., Gablonski, T. C., & Rabung, S. (2018). Prevalence of personality disorders in the general adult population in Western countries: systematic review and meta-analysis. *The British Journal of Psychiatry*, *213*(6), 709-715. https://doi.org/10.1192/bjp.2018.202

Wagner, T., Fydrich, T., Stiglmayr, C., Marschall, P., Salize, H. J., Renneberg, B., ... & Roepke, S. (2014). Societal cost-of-illness in patients with borderline personality disorder one year before, during and after dialectical behavior therapy in routine outpatient care. *Behaviour Research and Therapy*, *61*, 12-22. https://doi.org/10.1016/j.brat.2014.07.004

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Widiger, T. A., & Trull, T. J. (1998). Performance characteristics of the DSM-IIIR personality disorder criteria sets. In: Widiger, T. A., Frances, A. J., Pincus, H. A., Ross, R., First, M. B., Davis, W., & Kline, M. (Eds.). DSM-IV sourcebook, Vol. 4 (pp. 357-73). Arlington, VA, US: American Psychiatric Publishing

Wingenfeld, K., Spitzer, C., Mensebach, C., Grabe, H. J., Hill, A., Gast, U., ... & Driessen, M. (2010). The German Version of the Childhood Trauma Questionnaire (CTQ): Preliminary Psychometric Properties. *PPmP-Psychotherapie· Psychosomatik· Medizinische Psychologie*, *60*(11), 442-450. https://doi.org/10.1055/s-0030-1247564

Zanarini, M. C., Frankenburg, F. R., Hennen, J., Reich, D. B., & Silk, K. R. (2005). Psychosocial functioning of borderline patients and axis II comparison subjects followed prospectively for six years. *Journal of Personality Disorders*, *19*(1), 19-29. https://doi.org/10.1521/pedi.19.1.19.62178

Zettl. M., Volkert, J., Vögele, C., Herpertz, S. C., Kubera, K., & Taubner, S. (2019). Mentalization and Criterion A of the Alternative Model for Personality Disorders: Results from a clinical and nonclinical sample. *Personality Disorders: Theory, Research, & Treatment*. Manuscript in press.

Zimmerman, M. (1994). Diagnosing personality disorders: A review of issues and research methods. *Archives of General Psychiatry*, *51*(3), 225-245. https://doi.org/10.1001/archpsyc.1994.03950030061006

Zimmerman, M., & Mattia, J. I. (1999). Psychiatric diagnosis in clinical practice: is comorbidity being missed?. *Comprehensive Psychiatry*, *40*(3), 182-191. https://doi.org/10.1176/ajp.156.10.1570

Zimmermann, J., Altenstein, D., Krieger, T., Holtforth, M. G., Pretsch, J., Alexopoulos, J., ... & Leising, D. (2014). The structure and correlates of self-reported DSM-5 maladaptive personality traits: Findings from two German-speaking samples. *Journal of Personality Disorders*, *28*(4), 518-540. https://doi.org/10.1521/pedi_2014_28_130

Zimmermann, J. & Kerber, A., Rek, K., Hopwood C., & Krueger, R. (2019). A Brief but Comprehensive Review of Research on the Alternative DSM-5 Model for Personality Disorders. *Current Psychiatry Reports*, *21*(9). https://doi.org/10.1007/s11920-019-1079-z

Study 5

VI. Evers, O., **Schröder-Pfeifer, P.**, Möller, H., & Taubner, S. (Under review) The Impact of Trainee Attributes and Training Variables on Competence Deterioration: Results from a Longitudinal Study in Naturalistic German Psychotherapy Training. *Training and Education in Professional Psychology*. IF: 1.028

Declaration of author contributions: Evers, O.: Conceptualization, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project Administration, **Schröder-Pfeifer, P.:** Conceptualization, Writing – Original Draft, Writing – Review & Editing, Methodology, Software, Formal analysis, Visualization **Möller, H.:** Conceptualization, Funding Acquisition, Supervision **Svenja Taubner:** Conceptualization, Funding Acquisition, Project Administration, Writing – Review & Editing, Supervision

The Impact of Trainee Attributes and Training Variables on Competence Deterioration: Results from a Longitudinal Study in Naturalistic German Psychotherapy Training

Oliver Evers[1,2], Paul Schröder-Pfeifer[1,2], Heidi Möller[3], Svenja Taubner[1]

Affiliations:
[1]Institute for Psychosocial Prevention, Heidelberg University Hospital, Ruprecht Karl University of Heidelberg, Germany
[2]Psychological Institute, Ruprecht Karl University of Heidelberg, Germany
[3]Department of Psychology, University of Kassel, Germany

Corresponding author:
Oliver Evers
Institute for Psychosocial Prevention
Heidelberg University Hospital
Ruprecht Karl University of Heidelberg
Bergheimer Str. 54
D-69115 Heidelberg
Germany
Email: oliver.evers@med.uni-heidelberg.de
Tel.: (+49) 6221 56 4771
Fax: (+49) 6221 56 4702

Brief biographies:

Oliver Evers, M.Sc. Psych., is a PhD Student at Heidelberg University and a research assistant at Heidelberg University Hospital, where he also provides treatments in parent-infant psychotherapy. He is currently in postgraduate training for cognitive behavioral therapy and systemic therapy. His research interests include the competence development of psychotherapy trainees as well as early childhood interventions and developmental psychopathology in families at psychosocial risk.

Paul Schröder-Pfeifer, M.Sc. Psych., is a PhD Student at Heidelberg University and works as a research assistant at Heidelberg University Hospital. He is currently in training for psychodynamic therapy. He focuses on machine learning and missing data analysis and his research interests include mentalizing and epistemic trust.

Heidi Möller, PhD, is full professor of theory and methodology of counselling at the University of Kassel, department of psychology and head of the postgraduate program in coaching, organizational consulting, and supervision (COS). She is a psychologist, psychoanalyst, consultant, supervisor, and coach. She has worked as a consultant for commercial companies, public service, health services, and psychosocial services for the past 25 years.

Svenja Taubner, PhD, is full professor for psychosocial prevention at the University Hospital of Heidelberg, Germany, where she directs the institute for psychosocial prevention. She received her doctorate from the University of Bremen, Germany, and was full professor for clinical psychology and psychotherapy research at the University of Klagenfurt, Austria. She is Past President of the European Chapter of the Society for Psychotherapy Research and member of the national scientific psychotherapy chamber in Germany. Her research and professional interests include mentalization, development of psychological treatment for conduct disorder, and professional development in health service and psychotherapy training.

Abstract

Objectives: Competence deterioration during psychotherapy training can have a detrimental impact on treatment quality and trainee wellbeing. This study aimed to identify trainee and training attributes that predict deterioration. **Methods:** This study is an exploratory, secondary analysis of data from a 3-year longitudinal study on competence development in naturalistic psychotherapy training. The sample included 184 trainees for pre-assessment, of which 130 completed the post-assessment. Outcome was assessed for professional competence (knowledge, case-formulation competence, Healing Involvement, Stressful Involvement), personal competence (attributional complexity, introject affiliation), and relational competence (relatedness in client-treatments). A random forest algorithm was applied to predict competence deterioration from a broad set of predictor variables that included trainee sociodemographics, personality, attachment, childhood trauma, life satisfaction, therapeutic attitude, training elements, and training context. **Results:** Overall, 54% of trainees deteriorated in at least one outcome. Most important for prediction were variables from the domains of trainee attachment, life satisfaction, personality, therapeutic attitude, and childhood trauma. Training variables contributed little to prediction, except for personal therapy duration, which had an important protective effect. Sociodemographics did not predict deterioration. **Conclusions:** Trainee attributes have a high impact on competence deterioration. Training variables contribute little to prediction, except for personal therapy duration. Long-term personal therapies appear to be protective against competence deterioration. **Keywords:** psychotherapy training, therapist characteristics, professional development, negative effects, attachment

Public Significance: Our study found that competence decline during psychotherapy training is highly affected by personal attributes of trainees, like attachment, personal resources, and personality. Apart from personal therapy, few training elements seemed to be protective against competence decline. Our findings speak for establishing continuous support structures that help trainees reflect and cope with challenges over the whole duration of training.

The Impact of Trainee Attributes and Training Variables on Competence Deterioration: Results from a Longitudinal Study in Naturalistic German Psychotherapy Training

Competence problems among psychotherapy trainees are considered highly problematic for training and treatment quality (Nodop & Strauß, 2013) and previous research has addressed how to deal with problems once they become apparent (Vacha-Haase et al., 2019). However, few studies have quantified how many trainees experience meaningful competence deterioration during training (Dennhag & Ybrandt, 2013; Liness et al., 2019) and none of these studies investigated predictors of negative development. While a number of possible trainee or training attributes might be related to deterioration, there is no knowledge of empirically relevant predictors of deterioration.

The current study is a secondary analysis of data from a multidimensional outcome study [citation blinded for peer review]. In the original study, we investigated the competence development of 184 German psychotherapy trainees over three years and found an overall increase of professional and relational competence as well as stagnation of personal competence. Reliable change indices revealed that 54% of trainees deteriorated in at least one of the nine outcomes (ibid.). In the current investigation, we aim to use an exploratory analysis of auxiliary variables to predict which trainees deteriorated in any outcome during training. We will include a broad set of predictors, representing trainee attributes as well as training aspects.

Defining Deterioration in Psychotherapy Training

There is no agreed-upon definition of deterioration in psychotherapy training and deterioration has to be differentiated from challenging experiences. While trainees have described frequent challenges in training, such as self-criticism, anxiety, and stress (Murphy et al., 2018; Wilson et al., 2016), these negative experiences are not necessarily signs of competence deterioration. Qualitative research even indicates that a certain degree of challenges can be part of professional growth (Rønnestad & Skovholt, 2013). At the same time, it is possible that a decline in trainee competence, like conceptual competence, is not accompanied by negative experiences. Thus, in this study, deterioration will not be measured via emotional distress but via competence metrics, based

on a multidimensional competence profile (Bundespsychotherapeutenkammer [BPTK], 2008). Within this framework, deterioration will be defined as a meaningful decline in trainee competence.

The meaningfulness of trainees' competence decline can be measured through several indicators. The majority of training studies tend to report change on the group level (Rakovshik & McManus, 2010), but average effect sizes can mask substantial deterioration in subgroups of participants (Bauer et al., 2004). In contrast, cut-offs based on competency benchmarks (Fouad et al., 2009), are useful in measuring individual trainee progress by evaluating whether trainees fall below a certain threshold; but they might lead to over- or underestimating trainee deterioration, depending on how close trainees were to passing a threshold to begin with. This issue can be addressed through calculating Reliable Change Indices (RCI; Jacobson & Truax, 1991), which offer a way to evaluate whether an individual's competence decline can be considered reliable, based on sample variance and measure reliability. In line with recommendations for psychotherapy outcome research, this study will use RCIs as indicators for reliable deterioration (Bauer et al., 2004) and cut-off values where calculating an RCI is not possible.

Assessing and Predicting Deterioration

In competency-based education, relevant outcomes for psychotherapy training are defined through competence profiles, which are generally proposed by expert committees (Kaslow et al., 2004). This study is based on a profile of core competences for psychotherapists that was published by the Federal Chamber of Psychotherapists in Germany (BPTK, 2008). It was developed to guide a competency-oriented psychotherapy-training reform that came into effect in 2020. According to the profile, core-competences for psychotherapeutic licensure include professional-conceptual competence (e.g. knowledge, diagnostics, case conceptualization, technical skills), personal competence (e.g. self-reflection, emotional stability, coherent self, self-regulation), and relational competence (e.g. interpersonal competence and behaviors in the therapeutic relationship).

Since the competence profile (BPTK, 2008) does not specify competence measurements, outcomes in this study rely on instruments that are considered relevant for training and treatment

quality. *Professional-conceptual competence* will be assessed through a standardized knowledge-exam, ratings of trainees' case-conceptualization competence (Eells et al., 1998) and therapist work involvement (Orlinsky & Rønnestad, 2005). Therapist work involvement combines several aspects of trainees' work with client into two scales of global professional development, Healing Involvement and Stressful Involvement (Orlinsky & Rønnestad, 2005). *Personal competence* will be measured through assessments of self-reflection (Fletcher et al., 1986) and introject affiliation (Benjamin, 1995). Self-reflection will be operationalized as attributional complexity, which assesses trainees' interest in metacognition and the complexity of attributional schemata to human behavior (Fletcher et al., 1986). Introject affiliation describes the degree of affiliation/hostility in self-directed behavior (Benjamin, 1995). *Relational competence* will be assessed as self-reported relatedness in patient treatments, i.e. the degree of affiliation/hostility in patient-directed behavior (Benjamin, 1995). A detailed description of the constructs can be found in [blinded for peer review].

Deterioration in the aforementioned competences could pose a risk for treatment quality, patient safety, and trainee wellbeing. Among the measured outcomes, case-conceptualization competence, therapist work involvement, introject affiliation, and relatedness have predicted the quality of treatment process (Henry et al., 1990; Nissen-Lie et al., 2010) and outcome (Bruck et al., 2006; Easden & Fletcher, 2020; Nissen-Lie et al., 2017). Conceptually, deficits in self-reflective ability, low introject affiliation, and low relatedness in treatments could also lead to negative complementarity in patient-interactions (Henry et al., 1990). These negative interactions do not only pose risks for treatment quality and patient safety (Ackerman & Hilsenroth, 2001) but could ultimately lead to lower self-efficacy (Taubner et al., 2013), low work satisfaction (Orlinsky & Rønnestad, 2005), and work-related stress (Grundmann et al., 2013). Consequently, the measures in this study are potentially important indicators for adverse development during training.

Qualitative training studies imply that deterioration might be influenced by a variety of, possibly interrelated, trainee or training variables. Among training elements, studies have most often linked adverse trainee experiences to supervision (Wilson et al., 2016), personal therapy

(Murphy et al., 2018) or client treatments (Hill et al., 2007; Rønnestad & Skovholt, 2013) but it is still unknown whether training elements might cause deterioration or could serve as protective factors. Several personal attributes such as personality traits (DeNeve & Cooper, 1998), attachment (Allen et al., 2007) or biographical experiences (Anda et al., 2006) have been shown to positively and negatively influence wellbeing in the general population but it is unclear whether they also affect trainee development. Likewise, trainee's professional attributes like professional background (Rønnestad & Skovholt, 2013) or their therapeutic attitude (Sandell et al., 2004) might influence how they experience practical work with patients. Many of these variables are interrelated (Noftle & Shaver, 2006; Rizq & Target, 2010) and investigating a small number of predictors might lead to confounding sources of influence. This warrants an exploratory research approach in order to inform further targeted investigations in this field.

The Current study

The goal of this study is to identify possible predictors of competence deterioration over three years of psychotherapy training. The analysis will be based on a longitudinal outcome study of 184 German psychotherapy trainees, covering outcomes on three competence domains [blinded for peer review]. *Professional competence* is assessed via a knowledge exam, quality-ratings of trainees' case formulations (Eells et al., 1998), and therapist work involvement (Orlinsky & Rønnestad, 2005). *Personal competence* is measured via the attributional complexity scale (Fletcher et al., 1986) and the introject affiliation scale of the Intrex questionnaire. *Relational competence* is investigated through self-reported relatedness in patient interactions using the Intrex questionnaire (Benjamin, 1995). The Intrex assesses affiliation/relatedness at trainees "best times" and their "worst times".

The original outcome study [blinded for peer review] found an increase in professional-conceptual and in relational competence, while personal competence stagnated. There were several group by time effects, showing that cognitive-behavioral trainees gained more in Healing Involvement, relatedness and introject affiliation at worst than psychodynamic trainees. Meanwhile, psychodynamic trainees had higher overall levels of Healing Involvement and attributional

complexity as well as lower scores on relatedness at best. The study also found that 54% of trainees deteriorated in at least one outcome. The current study aims to use auxiliary variables, collected during the outcome study, to predict which trainees deteriorate in at least one of the outcome measures. In an exploratory approach, we will use a broad set of possible predictors from the domains of trainee sociodemographics, personality, attachment, childhood trauma, life satisfaction, therapeutic attitude as well as training elements and training context. An ensemble-based machine learning technique will be used to explore, which variables best predict trainee deterioration.

Methods

Recruitment and Participants

We included trainees in state-licensed adult psychotherapy training programs. We contacted 29 programs in Germany and 17 programs (58.62%) agreed to cooperate. Reasons for declining participation were (i) wanting to avoid overburdening trainees and (ii) objections to the psychometric assessments. Participating programs were 2 cognitive-behavioral (CBT), 2 psychoanalytic (PA), and 1 psychodynamic (PD) programs as well as 12 training centers offering separate programs in several of these modalities. Trainees were invited to participate via the program administration. The original outcome study also included a control group (CG) of 35 psychologists who were not in training. The CG data was not used in the current study because we aim to investigate trainee deterioration with possible predictors that relate to trainee attributes and training variables.

Of 730 trainees who were enrolled in the programs, 184 trainees (25.21%) participated in the pre-assessment (T1). Because of data protection regulations, we couldn't contact non-participants to assess reasons for refusal. A total of 130 trainees participated in the post-assessment after 3 years (T2; 29.35% dropout). Participants' descriptive data can be found in table 1. Differences between study completers and dropouts were tested for all study variables using a Bonferroni-Holm adjustment for multiple testing. At T1 there were no significant differences between completers and

dropouts but one difference test reached marginal significance. CBT candidates were more likely to drop out than PD candidates ($\chi^2(2)=12.07$; $p=.002$).

Measures

Professional Competence

Knowledge Exam: The exam was based on the German licensure test which covers factual knowledge, including psychopathological models, diagnostics, and medicine. The necessary knowledge is defined by national guidelines; specific questions are designed by an expert committee of psychotherapists. In this study, 20 multiple choice questions, with 4 possible answer options each, were taken from previous exams. Correct items were summed up to form a total score.

Case Formulation Content Coding Method (CFCCM; Eells et al., 1998): The CFCCM is a rating system to evaluate content and quality of case formulations. Content is coded by segmenting the text into idea units and coding them for the occurrence of content categories. The elaboration of each content category is rated on a 6-point scale and combined to form an *elaboration* score. Additionally, the quality of the case formulation is rated on five 6-point scales (*precision of language*, *complexity*, *overall coherence*, *treatment plan elaboration*, *goodness of fit*). The CFCCM has exhibited good to excellent interrater reliability (Eells et al., 1998). In this study, trainees' case formulations were based on a standardized patient video to which trainees answered five open questions. Raters were trained and blind to study groups. The formulation elaboration score and the quality scores were summed up to form a total score. The interrater reliability (ICC=.68) and Cronbach's α (α=.66.) of the total score were moderate.

Therapist Work Involvement Scales (TWIS; Orlinsky & Rønnestad, 2005): The TWIS is a self-report questionnaire to measure global professional competence in work with patients. It was developed in a factor-analytic approach from a larger set of conceptually-derived items, using a large transnational sample (Orlinsky & Rønnestad, 2005). The items form two principal dimensions, *Healing Involvement* and *Stressful Involvement.* Healing Involvement encompasses therapists' basic relational skills, relational agency, relational manner, feelings of flow during psychotherapeutic work

and use of constructive coping strategies during difficulties. Stressful Involvement assesses difficulties in practice, in-session feelings of anxiety and boredom, as well as avoidant coping strategies. Cronbach's α was good to excellent in previous samples (α=.82-.93; Hartmann et al., 2015) and reached acceptable to good levels (α=.74-85) in the current sample.

Personal and Relational Competence

Attributional Complexity Scale (ACS; Fletcher et al., 1986): The ACS assesses the complexity of attributional schemata, including the interest in exploring differentiated explanations for human-behavior. The 28 items form a single scale that showed a good internal consistency and a good test-retest reliability in validation studies (Fletcher et al., 1986). The ACS score is unrelated to social desirability and correlates with performance on attributional complexity tasks (Fletcher et al., 1986). In the current study, the internal consistency was excellent (α=.91).

Intrex Questionnaire short form (Benjamin, 1995): The Intrex is a self-report measure based on the SASB cluster model (Benjamin et al., 2006) to rate interpersonal and self-directed actions. It proposes two cluster surfaces to classify interpersonal behavior, a *transitive* surface that represents actions directed towards others and an *intransitive* surface that represents reactive interpersonal behavior. The third surface (*introject*) describes internal actions, directed towards oneself. Each surface represents a circumplex that is arranged along two axes, *affiliation* (love/relatedness vs. attack/recoil) and *interdependence* (emancipation/separation vs. control/submission). Participants are asked to rate their behaviors during their best times and their worst times. The construct validity of the two-axial structure could be confirmed in previous studies (Benjamin et al., 2006). In this study, trainees reported interpersonal behavior in patient treatments from their perspective and patients' perspectives. The affiliation scores were calculated according to Pincus et al. (1998). Interpersonal affiliation was averaged across surfaces and perspectives to form a single score for relatedness (Pincus et al., 1999). Cronbach's alphas ranged from α=.70 to α=.90.

Predictors

Therapeutic Attitudes Scales (TASC-2) – trainee version (Sandell et al., 2008): The TASC-2 is a self-report instrument to assess basic assumptions and beliefs about psychotherapy. It has scales that assess therapeutic styles (*neutrality*, s*upportiveness*, s*elf-doubt*), basic assumptions (*irrationality*, a*rtistry*, *pessimism*) and curative factors (*adjustment*, *kindness*, *insight*). The trainee version also assesses training context, satisfaction with each aspect of training, and theoretical interest. In previous studies, the TASC-scales were found to discriminate between therapeutic orientation (Sandell et al., 2004). In the current study Cronbach's α ranged from α=.54 to α=.87, with the exception of the pessimism scale (α<.5).

Questions on Life Satisfaction (FLZM; Henrich & Herschbach, 2000): The FLZ^M assesses the satisfaction with eight areas of life, using two items per area. One item is used to assess the subjective importance of each area; a second item is used to report the satisfaction with that area, creating a weighted satisfaction index for each area. The internal consistency reached α=.82 in the validation study (Henrich & Herschbach, 2000) and α=.64 in the current study.

NEO Five Factor Inventory (NEO-FFI; Borkenau & Ostendorf, 1993): The NEO-FFI is one of the most widely used questionnaires to assess personality traits. It contains the scales: *neuroticism*, *extraversion*, *openness to experience*, *agreeableness*, and *conscientiousness*. In German validation samples, it showed internal consistencies ranging from α=.63 to α=.83 (Körner et al., 2002) and reached α=.72 to α=.86 in the current sample.

Experiences in Close Relationships Revised (ECR-RD; Ehrenthal et al., 2009): The ECR-RD is a self-report questionnaire for assessing adult attachment in close relationships. It was developed based on an item-response theory analysis on previous attachment measures. The ECR-RD contains the scales *attachment anxiety* and *attachment avoidance*. The construct validity has been demonstrated with regard to other attachment measures and the internal consistency was excellent in previous studies (α=.92; Ehrenthal et al., 2009). In the current sample, Cronbach's α was good to excellent (α=.83 - .92).

Childhood Trauma Questionnaire (CTQ; Bernstein et al., 2003): The CTQ is a retrospective questionnaire on experiences of childhood abuse and neglect. It contains scales about *physical*, *sexual* and *emotional abuse* as well as *physical* and *emotional neglect* and *minimization/denial* of experiences. The German version adds a scale on *inconsistency experiences*, scoring unpredictable parenting behavior. In the German validation study, all scales had a good internal consistency (α=.80 - .89) except for the scale "physical neglect" (Klinitzke et al., 2012). In the current study Cronbach's α was acceptable to good (α=.78 - .87), except for "physical neglect" (α=.18).

Procedure

The study was designed as a naturalistic investigation with a pre measurement at the beginning of the study (T1) and a post measurement after three years of training (T2). Questionnaires were completed online. Knowledge exams and case formulations were completed in a supervised setting, in person at T1 and online at T2. Trainees were given 30 minutes each for the knowledge exam and case formulations. The study was approved by the ethics committee of [blinded for peer review]. All participants gave their informed consent for participating in the study.

At the time of the study[1], German Psychotherapy training was organized as post-graduate specialty training. The entry-level requirement was a 5-year academic degree in psychology. Contrary to public university education, psychotherapy training required high tuition fees. The training duration was 4200h over a minimum of three to five years. Required training elements were didactic instruction, two clinical internships, personal therapy/self-experience, and outpatient treatments under supervision. Licenses were obtained through a written and oral licensing exam. Trainees in this study were enrolled in CBT, PD or PA programs, which were the only psychotherapies financed by the public health insurance at the time.

[1] New laws and regulations came into effect in September 2020. In the future, universities will offer graduate programs in psychotherapy, followed by five years of post-licensing specialty training.

Statistical Analysis

The data contained 16.2% missing values. We imputed the missing data under the missing at random assumption. We utilized multiple imputations by chained equations (MICE) implemented in the MICE R package (version 3.6.0; van Buuren & Groothuis-Oudshoorn, 2011) with 60 iterations.

The goal of this study was to predict which trainees deteriorated in at least one competence outcome. In order to quantify deterioration among trainees, we computed reliable change indices (Jacobson & Truax, 1991), utilizing the measures' Cronbach's α in the computation of the standard error of the measurement, for all measures except for the knowledge exam. The knowledge test was not designed to form scales and thus Cronbach's α could not be computed. Instead, deterioration was measured via the official cut-off values for German licensing exams (60% correct answers). Deterioration for knowledge scores was defined as changing from a passing grade at T1 to failing at T2. To facilitate interpretation of results, predictor variables were organized in nine domains, namely training attributes (training aspects and training context) and trainee characteristics (personality traits, attachment strategies, life satisfaction, childhood trauma, therapeutic attitude, sociodemographics). Table 3 in the online supplement contains a full list of the predictor variables.

In order to predict competence deterioration, we employed a random forest (RF) algorithm. RF is a machine-learning technique, able to handle a high number of variables, even if they are correlated. It enables variable importance statistics that are more robust than commonly used linear model methods. RF is used to explore, which variables best divide the sample into two groups, i.e. trainees who deteriorated and trainees who did not. A number of classification trees are formed, to see, which predictor variables best divide the observations according to the outcome criterion. This process is repeated several times for random subsets of variables. In the end, all resulting classification trees are averaged, leading to an overall classification of each participant. The algorithm based classification can be compared to the actual outcome in the study in order to evaluate the classification accuracy. Each variable is assigned an importance score from 0 to 100, indicating how well it divides the observations.

Conditional inference RF was used in this study (Hothorn et al., 2006). Recursive feature elimination function "rfe" was used to identify the most stable predictors. Since the sample was too small to split it into training-, validation-, and test-set, we employed 5-fold cross-validation with 10 repeats to assess generalizability. We did not optimize hyperparameters to avoid overfitting through a spill-over of information. Gini impurity was chosen as a metric to find the optimal splits for each tree. Accuracy was used as classification metric. We calculated marginal effects using the tree traversal method. Mean importance was calculated for each variable domain. Statistical analyses were performed with R version 3.6.1 and the R package "caret" version 6.0-84 (Kuhn, 2008).

Results

Descriptive Statistics

Descriptive data for outcome variables and predictors is presented in table 2 and in table 3 of the online supplement respectively. In total, 70 trainees (53.85%) deteriorated; 46 trainees (35.38%) deteriorated in one outcome, 18 (13.45%) in two outcomes, and 6 trainees (4.62%) in more than two outcomes. Table 2 shows the deterioration rates per outcome. Deterioration rates were high for introject affiliation at worst (28.46%) and elevated for attributional complexity (10.77%), knowledge (8.46%), Stressful Involvement (8.46%), introject affiliation at best (7.69%), and case-conceptualization competence (6.15%).

Predicting Deterioration

In predicting which trainees deteriorated in at least one outcome, an accuracy of 66.85% was achieved at a no information rate of 53.85%. The no information rate is the observed rate of the more prevalent category and serves as a benchmark for the significance of our classification accuracy. We achieved a specificity of 58.66% and a sensitivity of 73.84%.

The final model includes 17 variables that were identified as the most stable predictors. Table 4 in the online supplement lists the individual importance values. Figure 1 shows the mean importance for each predictor category. Variables from the attachment domain (attachment avoidance, attachment anxiety) had a large predictive value. High average importance was also

found for life satisfaction (partner relationship, health), personality (extraversion, openness), therapeutic attitude (neutrality, insight) and childhood trauma (inconsistency experiences). Training aspects (current clinical internship at T1, observational learning, satisfaction with supervision) and training context (training program, part time training) were stable predictor variables but had small predictive values. An exception was the duration of personal therapy, which was among the most important predictor variables. All other variables did not represent stable predictors of deterioration and were excluded from the final model, among them all variables from the sociodemographic domain.

Figures 2-7 in the online supplement show partial dependency plots of the most important variable per category. The plots display the individual contribution of each variable to the chance for deterioration. The figures indicate a threshold effect; i.e. the likelihood to be classified in the deterioration group remained stagnant or showed a small change until a cut-off value, where the likelihood changed sharply. For attachment avoidance, high to very high values were associated with a higher chance of deterioration (figure 2). Longer personal therapies (figure 3) and a very high satisfaction with trainees' partner relationships (figure 4) led to a lower likelihood to deteriorate. Extraversion (figure 5) and neutrality (figure 6) followed similar patterns, while inconsistency experiences in childhood (figure 7) showed a more linear incline in deterioration likelihood. Among these variables, attachment avoidance, personal therapy duration, and relationship satisfaction were most influential in predicting deterioration.

Discussion

This study found that trainee variables were highly important in predicting competence deterioration, while training variables contributed little to prediction. Deterioration rates were highest for personal competence outcomes, namely introject affiliation at worst and attributional complexity. There was less, albeit still substantial, deterioration in professional competence. The most important predictor domain was trainee attachment, followed by variables from the domains life satisfaction, personality, therapeutic attitude, and childhood trauma (figure 1). Attachment

avoidance and trainee satisfaction with their partner relationship had the highest impact on deterioration likelihood. Training aspects and context played a comparatively minor role, with the exception of personal therapy duration. Sociodemographic variables had no predictive value.

The overall predictor pattern in this study could speak for an important role of trainee-related risk factors and protective factors that impact competence deterioration. Personal trainee variables were most important in predicting deterioration (figure 1) and overwhelmingly came into play once trainees passed a certain cut-off value (figures 2-7). For instance, the chance of deterioration increased sharply at attachment avoidance levels, typically found in clinical samples (Ehrenthal et al., 2009). This pattern implies that certain trainee attributes do not necessarily represent general vulnerabilities, but rather take the form of risk factors (attachment avoidance, childhood trauma, extraversion) and protective factors (satisfaction in partner relationship, neutral therapeutic style) above certain thresholds (cf. Nodop & Strauß, 2013). Mechanisms, through which personal variables impact trainee development, are highly speculative and might involve the influence of attachment patterns and childhood trauma on trainee wellbeing (Allen et al., 2007; Anda et al., 2006), their impact on interpersonally challenging client treatments (Schauenburg et al., 2010), and maladaptive reactions through highly expressed or rigid personality structures.

Surprisingly, most training variables had little to no importance for prediction, with the exception of personal therapy duration. Years spent in personal therapy was a protective factor, independent of dosage, which points to a buffering effect of support structures (Rønnestad & Skovholt, 2013) that extended over the whole duration of training. Program orientation did not contribute to predicting deterioration, despite our earlier findings that overall competence development differed between PD and CBT trainees [blinded for peer review]. These disparate findings could be due to including a number of possible covariates that may vary between orientations, like therapeutic attitude and personality (Taubner et al., 2014). The finding that satisfaction with supervision contributed little to predicting deterioration, while frequency did not contribute at all, could show that supervision quality might be more protective than the amount of

supervision. In that vein, satisfaction might still be a poor indicator of supervision quality and measures like the supervisory alliance might have yielded better predictions (Falender et al., 2014).

The findings that personal competence outcomes were highly affected by deterioration and mostly predicted by personal attributes give rise to the question, whether we captured negative personal development that was unrelated to training. On the one hand, personal competence outcomes are not context-specific and could possibly be affected by trainees' private lives (Benjamin et al., 2006; Fletcher et al., 1986). On the other hand, our findings indicate that personal relationship crises were unlikely causes of competence deterioration, because the likelihood to deteriorate was stable across the whole range of trainees' partnership satisfaction, except for extremely satisfied trainees. Ultimately, this study can neither confirm nor rule out the presence of personal crises. Nevertheless, the findings point to personal trainee vulnerabilities that might not be sufficiently addressed in training.

Limitations

This study used a data-driven, exploratory approach which calls for a cautious interpretation of our findings. Specifically, due to the sample size, we utilized cross-validation instead of separate test-sets to assess the generalizability of our results and the findings need to be validated in future studies. In order to explore the process of competence deterioration as a whole, we combined all trainees who deteriorated in any outcome into one group, but the underlying processes might differ according to outcome measure. Due to the extensive study design, we couldn't use observational ratings for relational competence, which might have yielded more instances of deterioration on that domain. Conceptually, the competence profile did not fully specify empirically measureable constructs and our measurements had to be inferred from content descriptions, which highlights the importance of defining empirically-informed core-competences for licensure (Kaslow et al., 2009).

Implications for Research and Training

Our results underline that training studies should consistently assess and report adverse events. We designed this secondary analysis to identify target variables for future investigations and

found that the impact of personal trainee variables on negative processes should be more closely examined, especially with regard to attachment, personality, and childhood trauma. These variables could be studied as moderators of competence development and research could focus on whether they affect, how trainees process training challenges. Methodically, our research approach shows that simple machine-learning algorithms can be useful exploratory tools in research fields where there is little previous knowledge.

Due to the exploratory nature of the study, practical implications should be viewed with extreme caution. Nevertheless, our results could indicate an important role of personal therapy in protecting against competence deterioration. The findings speak for emphasizing continuous, long-term support during training, especially in the face of trainee risk factors. The high extend of deterioration, together with the unsubstantial impact of most training variables, highlight a need for monitoring competence deterioration during training and tailoring training methods to better address negative processes. While the results could hypothetically be used to inform trainee selection, basing selection processes on personal vulnerabilities would most likely lead to trainees not disclosing vulnerabilities in the first place.

Conclusion

The current study found that trainee attributes, specifically personal trainee variables, were most important in predicting competence deterioration over three years of training. Training variables, except for the duration of personal therapy, contributed little to prediction. These findings show that training research should consistently take adverse events and negative development during training into account and should also consider personal variables that contribute to these effects. Training programs could implement long-term support over the whole duration of training, e.g. in the form of personal therapy, and should devise approaches that help address negative development during training.

References

Ackerman, S. J., & Hilsenroth, M. J. (2001). A review of therapist characteristics and techniques negatively impacting the therapeutic alliance. *PSYCHOTHER-THEOR RES*, *38*(2), 171-185. https://doi.org/10.1037/0033-3204.38.2.171

Allen, J. P., Porter, M., McFarland, C., McElhaney, K. B., & Marsh, P. (2007). The relation of attachment security to adolescents' paternal and peer relationships, depression, and externalizing behavior. *CHILD DEV*, *78*(4), 1222-1239. https://doi.org/10.1111/j.1467-8624.2007.01062.x

Anda, R. F., Felitti, V. J., Bremner, J. D., Walker, J. D., Whitfield, C., Perry, B. D., Dube, S. R., & Giles, W. H. (2006). The enduring effects of abuse and related adverse experiences in childhood. *EUR ARCH PSY CLIN N*, *256*(3), 174-186. https://doi.org/10.1007/s00406-005-0624-4

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical Significance Methods: A Comparison of Statistical Techniques. *J PERS ASSESS*, *82*(1), 60-70. https://doi.org/10.1207/s15327752jpa8201_11

Benjamin, L. S. (1995). *SASB Intrex short form user's manual*. University of Utah.

Benjamin, L. S., Rothweiler, J. C., & Critchfield, K. L. (2006). The use of Structural Analysis of Social Behavior (SASB) as an assessment tool. *ANNU REV CLIN PSYCHO*, *2*(1), 83-109. https://doi.org/10.1146/annurev.clinpsy.2.022305.095337

Bernstein, D. P., Stein, J. A., & Newcomb, M. D. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *CHILD ABUSE NEGLECT*, *27*, 169-190.

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar:(NEO-FFI); nach Costa und McCrae*. Hogrefe.

Bruck, E., Winston, A., Aderholt, S., & Muran, J. C. (2006). Predictive Validity of Patient and Therapist Attachment and Introject Styles. *AM J PSYCHOTHER*, *60*(4), 393-406. https://doi.org/10.1176/appi.psychotherapy.2006.60.4.393

Bundespsychotherapeutenkammer. (2008). *Kernkompetenzen von Psychotherapeutinnen und Psychotherapeuten*.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin*, *124*(2), 197-229. https://doi.org/10.1037/0033-2909.124.2.197

Dennhag, I., & Ybrandt, H. (2013). Trainee psychotherapists' development in self-rated professional qualities in training. *PSYCHOTHER*, *50*(2), 158-166. https://doi.org/10.1037/a0033045

Easden, M. H., & Fletcher, R. B. (2020). Therapist competence in case conceptualization and outcome in CBT for depression. *PSYCHOTHER RES*, *30*(2), 151-169. https://doi.org/10.1080/10503307.2018.1540895

Eells, T. D., Kendjelic, E. M., & Lucas, C. P. (1998). What's in a case formulation? Development and use of a content coding manual. *J Psychother Pract Res*, *7*(2), 144-153.

Ehrenthal, J. C., Dinger, U., Lamla, A., Funken, B., & Schauenburg, H. (2009). Evaluation der deutschsprachigen Version des Bindungsfragebogens „Experiences in Close Relationships – Revised" (ECR-RD). *Psychother Psych Med*, *59*(06), 215-223. https://doi.org/10.1055/s-2008-1067425

Falender, C. A., Shafranske, E. P., & Ofek, A. (2014). Competent clinical supervision: Emerging effective practices. *Counselling Psychology Quarterly*, *27*(4), 393-408. https://doi.org/10.1080/09515070.2014.934785

Fletcher, G. J. O., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *J PERS SOC PSYCHOL*, *51*(4), 875-884. https://doi.org/10.1037/0022-3514.51.4.875

Fouad, N. A., Grus, C. L., Hatcher, R. L., Kaslow, N. J., Hutchings, P. S., Madson, M. B., Collins, F. L., Jr., & Crossman, R. E. (2009). Competency benchmarks: A model for understanding and measuring competence in professional psychology across training levels. *TRAIN EDUC PROF PSYC*, *3*(4, Suppl), S5-S26. https://doi.org/10.1037/a0015832

436 Grundmann, J., Sude, K., Löwe, B., & Wingenfeld, K. (2013). Arbeitsbezogene Stressbelastung und
437 psychische Gesundheit: Eine Befragung von Psychotherapeutinnen und -therapeuten in
438 Ausbildung. *Psychother Psych Med*, *63*(03/04), 145-149. https://doi.org/10.1055/s-0032-
439 1333292
440 Hartmann, A., Joos, A., Orlinsky, D. E., & Zeeck, A. (2015). Accuracy of therapist perceptions of
441 patients' alliance: Exploring the divergence. *PSYCHOTHER RES*, *25*(4), 408-419.
442 https://doi.org/10.1080/10503307.2014.927601
443 Henrich, G., & Herschbach, P. (2000). Questions on Life Satisfaction (FLZM) - A short questionnaire
444 for assessing subjective quality of life. *EUR J PSYCHOL ASSESS*, *16*(3), 150-159.
445 https://doi.org/10.1027//1015-5759.16.3.150
446 Henry, W. P., Schacht, T. E., & Strupp, H. H. (1990). Patient and therapist introject, interpersonal
447 process, and differential psychotherapy outcome. *J CONSULT CLIN PSYCH*, *58*(6), 768-774.
448 https://doi.org/10.1037/0022-006X.58.6.768
449 Hill, C. E., Sullivan, C., Knox, S., & Schlosser, L. Z. (2007). Becoming psychotherapists: Experiences of
450 novice trainees in a beginning graduate class. *PSYCHOTHER-THEOR RES*, *44*(4), 434-449.
451 https://doi.org/10.1037/0033-3204.44.4.434
452 Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference
453 Framework. *J COMPUT GRAPH STAT*, *15*(3), 651-674.
454 https://doi.org/10.1198/106186006X133933
455 Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful
456 change in psychotherapy research. *J CONSULT CLIN PSYCH*, *59*(1), 12-19.
457 https://doi.org/10.1037/0022-006X.59.1.12
458 Kaslow, N. J., Borden, K. A., Collins Jr., F. L., Forrest, L., Illfelder-Kaye, J., Nelson, P. D., Rallo, J. S.,
459 Vasquez, M. J. T., & Willmuth, M. E. (2004). Competencies Conference: Future Directions in
460 Education and Credentialing in Professional Psychology. *J CLIN PSYCHOL*, *60*(7), 699-712.
461 https://doi.org/10.1002/jclp.20016

462 Kaslow, N. J., Grus, C. L., Campbell, L. F., Fouad, N. A., Hatcher, R. L., & Rodolfa, E. R. (2009).
463 Competency Assessment Toolkit for professional psychology. *TRAIN EDUC PROF PSYC*, *3*(4),
464 S27-S45. https://doi.org/10.1037/a0015833
465 Klinitzke, G., Romppel, M., Häuser, W., Brähler, E., & Glaesmer, H. (2012). Die deutsche Version des
466 Childhood Trauma Questionnaire (CTQ) – psychometrische Eigenschaften in einer
467 bevölkerungsrepräsentativen Stichprobe. *Psychother Psych Med*, *62*(02), 47-51.
468 https://doi.org/10.1055/s-0031-1295495
469 Körner, A., Geyer, M., & Brähler, E. (2002). Das NEO-Fünf-Faktoren Inventar (NEO-FFI). *Diagnostica*,
470 *48*(1), 19-27. https://doi.org/10.1026//0012-1924.48.1.19
471 Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *J STAT SOFTW*, *28*(5).
472 https://doi.org/10.18637/jss.v028.i05
473 Liness, S., Beale, S., Lea, S., Byrne, S., Hirsch, C. R., & Clark, D. M. (2019). Multi-professional IAPT CBT
474 training: clinical competence and patient outcomes. *BEHAV COGN PSYCHOTH*, *47*(6), 672-
475 685. https://doi.org/10.1017/S1352465819000201
476 Murphy, D., Irfan, N., Barnett, H., Castledine, E., & Enescu, L. (2018). A systematic review and meta-
477 synthesis of qualitative research into mandatory personal psychotherapy during training.
478 *Counselling & Psychotherapy Research*, *18*(2), 199-214. https://doi.org/10.1002/capr.12162
479 Nissen-Lie, H. A., Monsen, J. T., & Rønnestad, M. H. (2010). Therapist predictors of early patient-
480 rated working alliance: A multilevel approach. *PSYCHOTHER RES*, *20*(6), 627-646.
481 https://doi.org/10.1080/10503307.2010.497633
482 Nissen-Lie, H. A., Rønnestad, M. H., Høglend, P. A., Havik, O. E., Solbakken, O. A., Stiles, T. C., &
483 Monsen, J. T. (2017). Love yourself as a person, doubt yourself as a therapist? *CLIN PSYCHOL*
484 *PSYCHOT*, *24*(1), 48-60. https://doi.org/10.1002/cpp.1977
485 Nodop, S., & Strauß, B. (2013). Mangelnde Eignung bei angehenden Psychotherapeuten.
486 *Psychotherapeut*, *58*(5), 446-454. https://doi.org/10.1007/s00278-013-1001-9

487 Noftle, E. E., & Shaver, P. R. (2006). Attachment dimensions and the big five personality traits:
488 Associations and comparative ability to predict relationship quality. *J RES PERS*, *40*(2), 179-
489 208. https://doi.org/10.1016/j.jrp.2004.11.003
490 Orlinsky, D. E., & Rønnestad, M. H. (Eds.). (2005). *How psychotherapists develop : a study of*
491 *therapeutic work and professional growth* (1st ed.). APA.
492 Pincus, A. L., Dickinson, K. A., Schut, A. J., Castonguay, L. G., & Bedics, J. (1999). Integrating
493 interpersonal assessment and adult attachment using SASB. *EUR J PSYCHOL ASSESS*, *15*(3),
494 206-220. https://doi.org/10.1027//1015-5759.15.3.206
495 Pincus, A. L., Newes, S. L., Dickinson, K. A., & Ruiz, M. A. (1998). A comparison of three indexes to
496 assess the dimensions of Structural Analysis of Social Behavior. *J PERS ASSESS*, *70*(1), 145-
497 170. https://doi.org/10.1207/s15327752jpa7001_10
498 Rakovshik, S. G., & McManus, F. (2010). Establishing evidence-based training in cognitive behavioral
499 therapy: A review of current empirical findings and theoretical guidance. *CLIN PSYCHOL REV*,
500 *30*(5), 496-516. https://doi.org/http://dx.doi.org/10.1016/j.cpr.2010.03.004
501 Rizq, R., & Target, M. (2010). 'If that's what I need, it could be what someone else needs.' Exploring
502 the role of attachment and reflective function in counselling psychologists' accounts of how
503 they use personal therapy in clinical practice: a mixed methods study. *BRIT J GUID COUNS*,
504 *38*(4), 459-481. https://doi.org/10.1080/03069885.2010.503699
505 Rønnestad, M. H., & Skovholt, T. M. (2013). *The developing practitioner : growth and stagnation of*
506 *therapists and counselors*. Routledge.
507 Sandell, R., Carlsson, J., Schubert, J., Broberg, J., Lazar, A., & Grant, J. (2004). Therapist attitudes and
508 patient outcomes: I. Development and validation of the Therapeutic Attitudes Scales (TASC-
509 2). *PSYCHOTHER RES*, *14*(4), 469-484. https://doi.org/10.1093/ptr/kph039
510 Sandell, R., Taubner, S., Rapp, A., Visbeck, A., & Kächele, H. (2008). *Psycho-Therapeutische Haltung*
511 *Ausbildungsversion (ThAt-AV)*. Universität Ulm.

Schauenburg, H., Buchheim, A., Beckh, K., Nolte, T., Brenk-Franz, K., Leichsenring, F., Strack, M., & Dinger, U. (2010). The influence of psychodynamically oriented therapists' attachment representations on outcome and alliance in inpatient psychotherapy. *Psychotherapy Research*, *20*(2), 193-202. https://doi.org/10.1080/10503300903204043

Taubner, S., Munder, T., Möller, H., Hanke, W., & Klasen, J. (2014). Selbstselektionsprozesse bei der Wahl des therapeutischen Ausbildungsverfahrens: Unterschiede in therapeutischen Haltungen, Persönlichkeitseigenschaften und dem Mentalisierungsinteresse. *Psychother Psych Med*, *64*(06), 214-223. https://doi.org/10.1055/s-0033-1358720

Taubner, S., Zimmermann, J., Kächele, H., Möller, H., & Sell, C. (2013). The relationship of introject affiliation and personal therapy to trainee self-efficacy: A longitudinal study among psychotherapy trainees. *PSYCHOTHER*, *50*(2), 167-177. https://doi.org/10.1037/a0029819

Vacha-Haase, T., Elman, N. S., Forrest, L., Kallaugher, J., Lease, S. H., Veilleux, J. C., & Kaslow, N. J. (2019). Remediation plans for trainees with problems of professional competence. *TRAIN EDUC PROF PSYC*, *13*(4), 239-246. https://doi.org/10.1037/tep0000221

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *J STAT SOFTW*, *1*(3). https://doi.org/10.18637/jss.v045.i03

Wilson, H. M. N., Davies, J. S., & Weatherhead, S. (2016). Trainee therapists' experiences of supervision during training: A meta-synthesis. *CLIN PSYCHOL PSYCHOT*, *23*(4), 340-351. https://doi.org/10.1002/cpp.1957

532 **Tables**

Table 1

Demographic data of participants (T1)

	Total (n=184)	Completers (n=130)
Age (*M* / *SD*)	31.42 (6.67)	33.48 (6.45)
Gender		
female	84.2%	86.9%
male	15.8%	13.1%
Semester (*M* / *SD*)	2.30 (1.82)	2.25 (1.81)
Orientation		
cognitive-behavioral	34.8%	26.9%
psychoanalytic	17.9%	20.0%
psychodynamic	47.3%	53.1%

533

Table 2

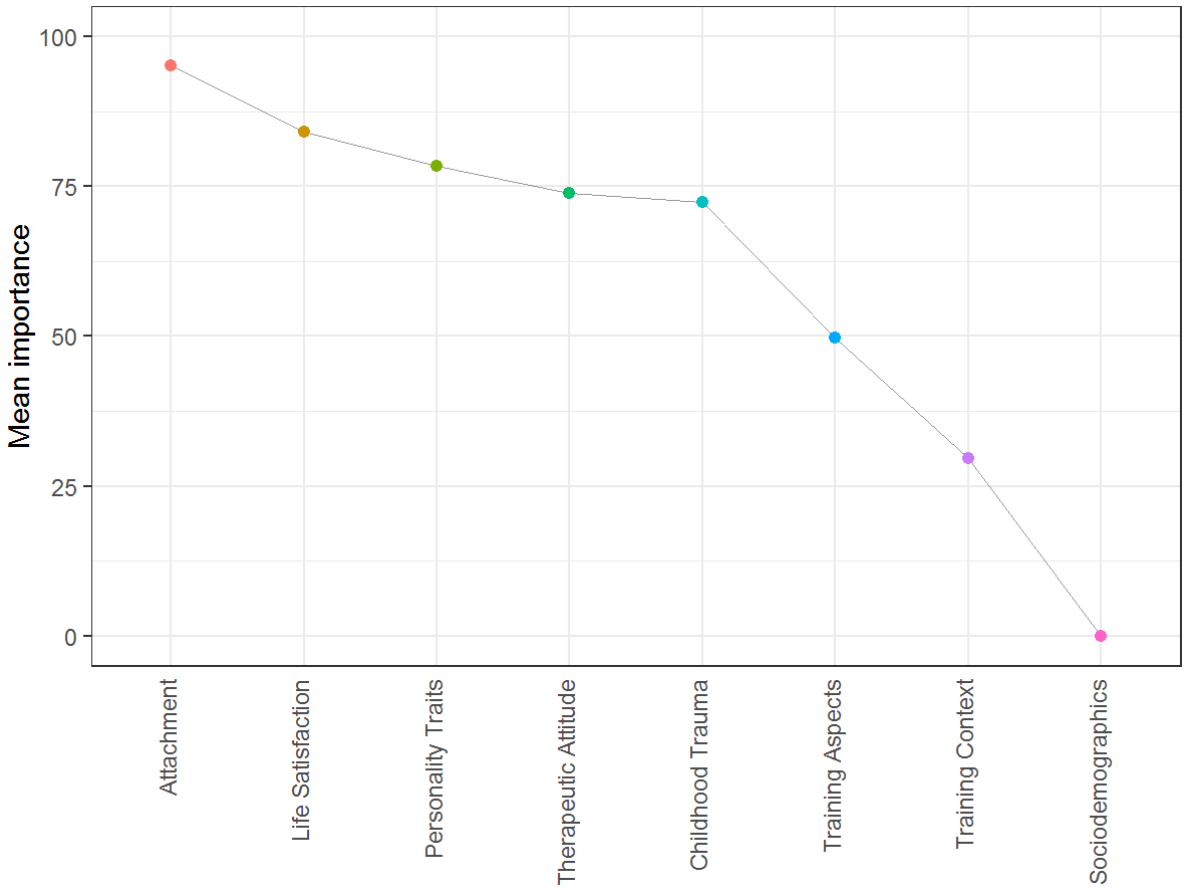
Descriptive data and deterioration rates of outcome variables

	T1		T2		Deterioration
Professional / conceptual	*M (SD)*	*range*	*M (SD)*	*range*	*%*
Knowledge	10.16 (2.43)	2.43 – 5.00	12.02 (2.75)	5.00 – 18.00	8.46
Case-Conceptualization	24.13 (5.17)	8.00 – 38.00	30.59 (6.01)	14.55 – 44.00	6.15
Healing Involvement	10.34 (1.16)	7.08 – 12.88	11.08 (1.04)	7.44 – 13.96	1.54
Stressful Involvement	4.74 (1.50)	1.45 – 9.23	4.82 (1.55)	0.95 – 10.77	8.46
Personal					
Attributional complexity	5.35 (0.76)	2.25 – 6.86	5.43 (0.73)	3.14 – 6.75	10.77
Introject affiliation at best	71.07 (20.00)	-3.30 – 100.80	78.24 (25.44)	-32.40 – 100.80	7.69
Introject affiliation at worst	23.22 (39.26)	-70.05 – 96.30	4.51 (44.88)	-100.80 – 100.80	28.46
Relational					
Relatedness at best	55.02 (15.41)	14.03 – 84.82	64.16 (16.16)	-11.7 – 100.80	3.85
Relatedness at worst	2.98 (29.81)	-84.23 – 63.00	10.56 (25.72)	-59.03 – 78.68	2.31

Note: *N*=130

Figure 1

Mean importance of variable domains, predicting systematic deterioration during training



Online Supplement

Tables

Table 3

Descriptive data on predictor variables

	M / %	*SD*	*Min*	*Max*
Attachment strategies				
Attachment anxiety (T1)	2.51	1.02	1.00	5.11
Attachment anxiety (T2)	2.52	1.08	1.06	6.00
Attachment anxiety change	0.10	0.86	-2.28	2.78
Attachment avoidance (T1)	2.19	0.82	1.00	6.11
Attachment avoidance (T2)	2.19	0.83	1.00	5.94
Attachment avoidance change	-0.07	0.88	-2.89	3.11
Childhood trauma				
Emotional abuse	9.62	3.74	5.00	21.00
Emotional neglect	10.85	3.82	5.00	24.00
Inconsistency experiences	7.06	3.10	3.00	15.00
Minimization/denial	0.11	0.49	0.00	3.00
Physical abuse	5.65	1.67	5.00	14.00
Physical neglect	7.03	2.05	5.00	13.00
Sexual abuse	5.91	2.13	5.00	17.00
Life satisfaction				
Family life/children	4.89	7.59	-12.00	20.00
Friends/acquaintances	8.08	6.56	-4.00	20.00
Global	6.00	3.66	-3.00	13.88
Health	7.69	6.71	-12.00	20.00
Housing/living conditions	5.61	5.67	-9.00	20.00
Income/financial security	2.39	6.4	-12.00	20.00
Leisure time/hobbies	5.48	6.45	-9.00	20.00
Occupation/work	7.04	6.54	-12.00	20.00
Partner relationship/sexuality	6.84	8.69	-12.00	20.00
Personality traits				
Agreeableness	35.37	5.07	19.00	48.00
Conscientiousness	31.55	5.64	16.00	43.00
Extraversion	30.61	5.31	17.00	44.00
Neuroticism	20.52	7.59	3.00	41.00
Openness	36.03	5.43	16.00	47.00
Sociodemographics				
Age	31.48	6.45	24.00	55.00
Additional therapy training(s)				
Yes	9.2%			
No	90.8%			
Additional university degree				
Yes	6.9%			

No	93.1%			
Civil Status				
single (unmarried)	72.3%			
married	23.8%			
divorced	3.8%			
Gender				
female	86.9%			
male	13.1%			
Relationship status				
in a relationship	79.2%			
not in a relationship	20.8%			
Therapeutic attitude				
Adjustment	1.95	0.49	0.46	3.15
Artistry	2.26	0.55	0.80	3.80
Insight	2.46	0.65	0.42	3.75
Irrationality	2.38	0.52	0.75	3.75
Kindness	3.07	0.50	1.80	4.00
Neutrality	2.24	0.43	1.18	3.27
Pessimism	1.71	0.38	0.60	2.60
Self-doubt	1.10	0.50	0.11	2.78
Supportiveness	2.50	0.43	1.00	3.30
Theoretical Breadth (T1)	4.75	1.49	0.00	8.00
Theoretical Breadth (T2)	4.62	1.33	2.00	8.00
Training aspects				
Clinical internship (T1)				
Currently in internship	60.8%			
Finished internship	12.3%			
Frequency of observational learning (T1)[1]	3.42	0.96	1.00	5.00
Frequency of observational learning (T2)[1]	3.38	1.12	1.00	5.00
Frequency of supervision (T1)				
regularly	20.0%			
occasionally	22.3%			
none	57.7%			
Frequency of supervision (T2)				
regularly	88.5%			
occasionally	6.9%			
none	4.6%			
General satisfaction with training (T1)[2]	3.14	0.63	1.00	4.00
General satisfaction with training (T2)[2]	3.89	0.72	2.00	5.00
Number of patients in the last year (T1)	6.08	11.74	0.00	80.00
Number of patients in the last year (T2)	24.51	32.23	0.00	220.00
Satisfaction with didactic seminars (T2) [2]	3.81	0.76	2.00	5.00
Satisfaction with personal therapy (T1)[2]	4.22	0.83	1.00	5.00
Satisfaction with personal therapy (T2)[2]	4.14	0.89	1.00	5.00
Satisfaction with supervision (T2)[2]	4.05	0.86	1.00	5.00
Total hours of personal therapy (T1)	135.61	143.71	0.00	710.00
Total hours of personal therapy (T2)	152.63	104.99	7.00	510.00
Total years of personal therapy (T1)	2.47	2.86	0.00	17.00

Total years of personal therapy (T2)	2.86	1.59	0.25	10.50
Type of personal therapy (T1)				
Mandatory training therapy / self-experience	52.3%			
Personal psychotherapy outside training	7.7%			
Voluntary training therapy / self-experience in different training	2.3%			
Combination of personal psychotherapy / training therapy / self-experience	10.0%			
Other	3.1%			
None	24.6%			
Training context				
Full time/part time training				
full time	60.0%			
part time	40.0%			
Orientation				
cognitive-behavioral	26.9%			
psychoanalytic	20.0%			
psychodynamic	53.1%			
Semester (T1)	2.25	1.81	1.00	12.00
Training temporarily interrupted				
yes	9.2%			
no	90.8%			

Note: [1]Rated on a scale from 1(never) to 5(often). [2]Rated on a scale of 1(not at all) to 5 (very).

542

543

Table 4

Importance of individual variables by category

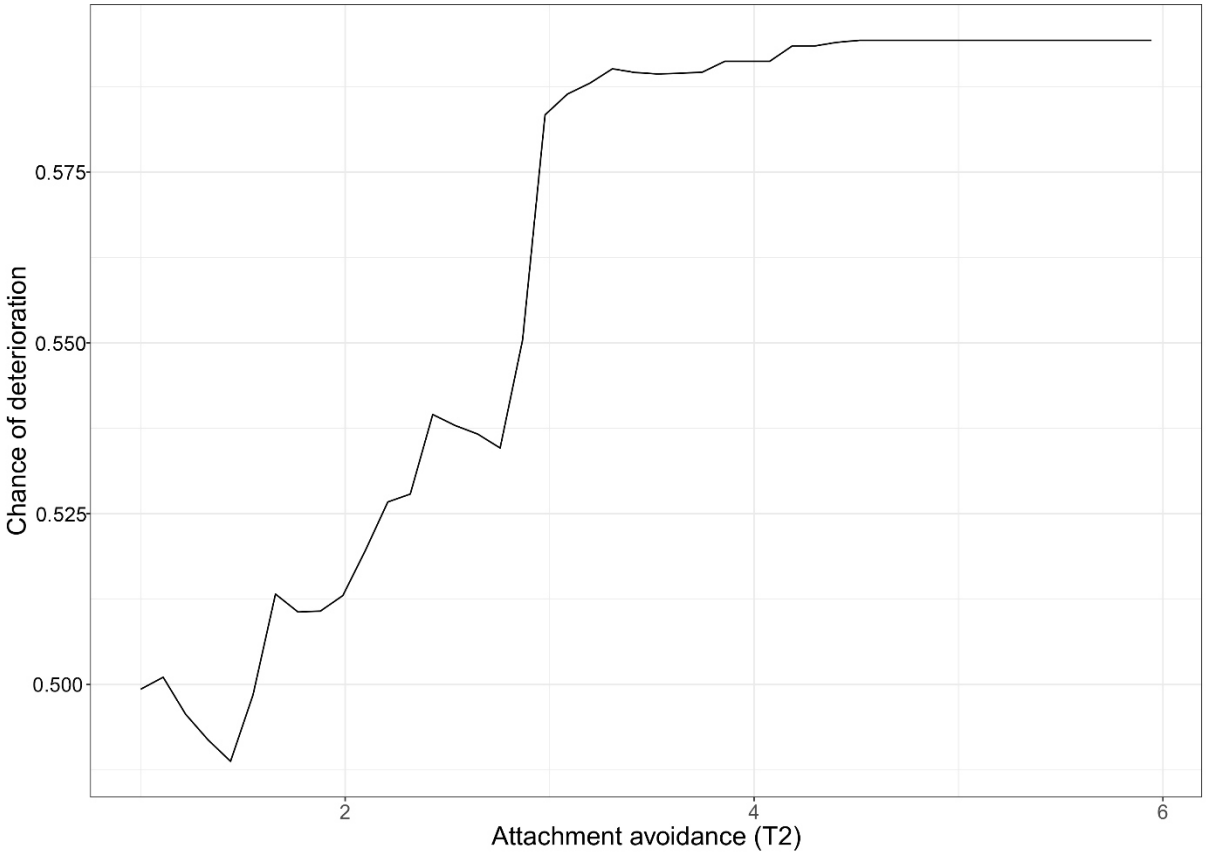
Category	Variable	Importance[1]
Attachment strategies	Attachment avoidance (T2)	100.00
	Attachment anxiety (T2)	98.35
	Attachment anxiety (T1)	86.85
Life satisfaction	Partner relationship/sexuality	89.40
	Health	78.78
Personality traits	Extraversion	83.45
	Openness	73.15
Therapeutic attitude	Neutrality	79.30
	Insight	68.44
Childhood trauma	Inconsistency experiences	72.35
Training aspects	Total years of personal therapy (T2)	89.95
	Clinical internship (T1) – currently	53.28
	Frequency of observational learning (T2)	43.11
	Frequency of observational learning (T1)	35.05
	Satisfaction with supervision (T2)	27.59
Training context	Training program	16.92
	Full time/part time	12.25

Note: The table shows predictor variables in the final model. Variables are the most stable predictors, identified through recursive feature elimination.

[1]Importance values range from 0 – 100, 100 being most important for prediction.

Figure 2

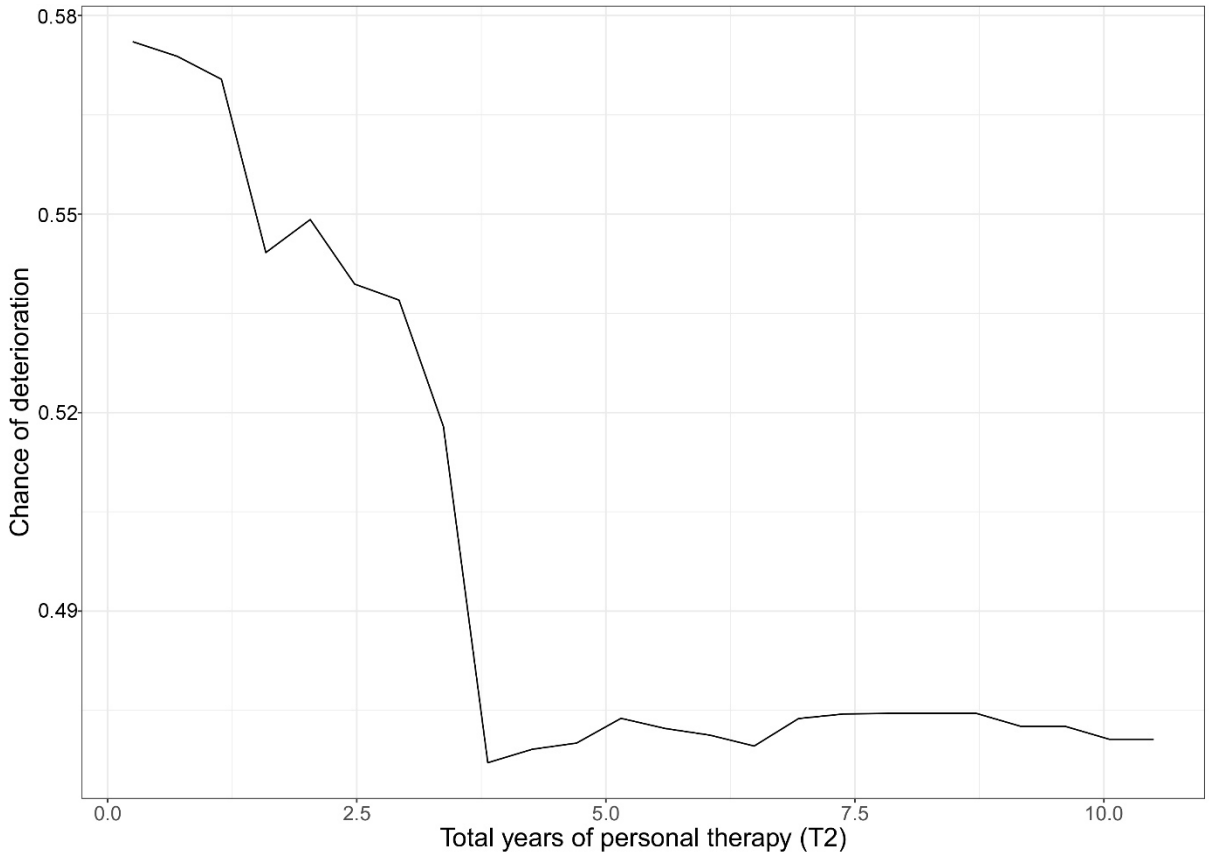
Partial dependency plot for attachment avoidance (T2)



Note. Individual contribution of attachment avoidance to the chance of deterioration. Scores of attachment avoidance represent assessments of adult attachment via self-reported attachment strategies in close personal relationships. Higher values on the y-axis indicate a higher likelihood to deteriorate.

Figure 3

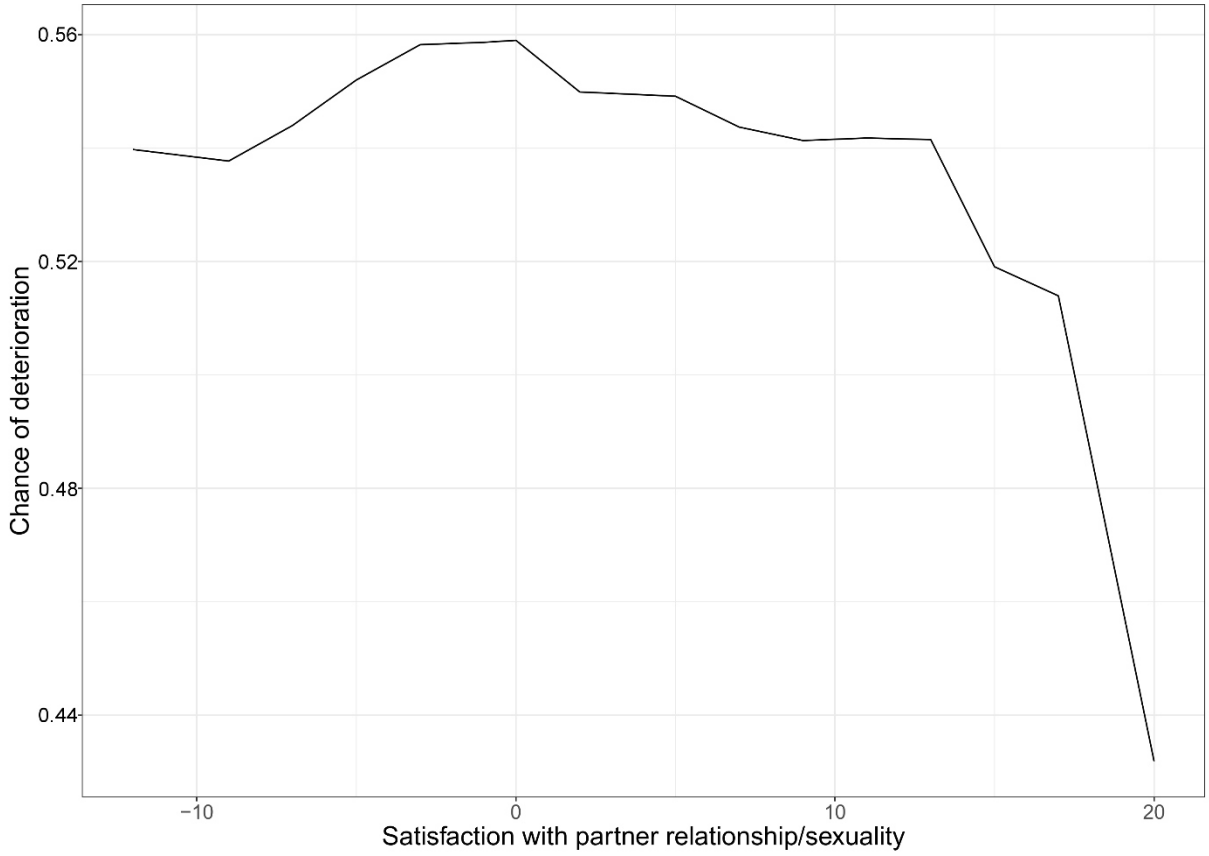
Partial dependency plot for personal therapy duration



Note. Individual contribution of personal therapy duration to the chance of deterioration. Higher values on the y-axis indicate a higher likelihood to deteriorate.

Figure 4

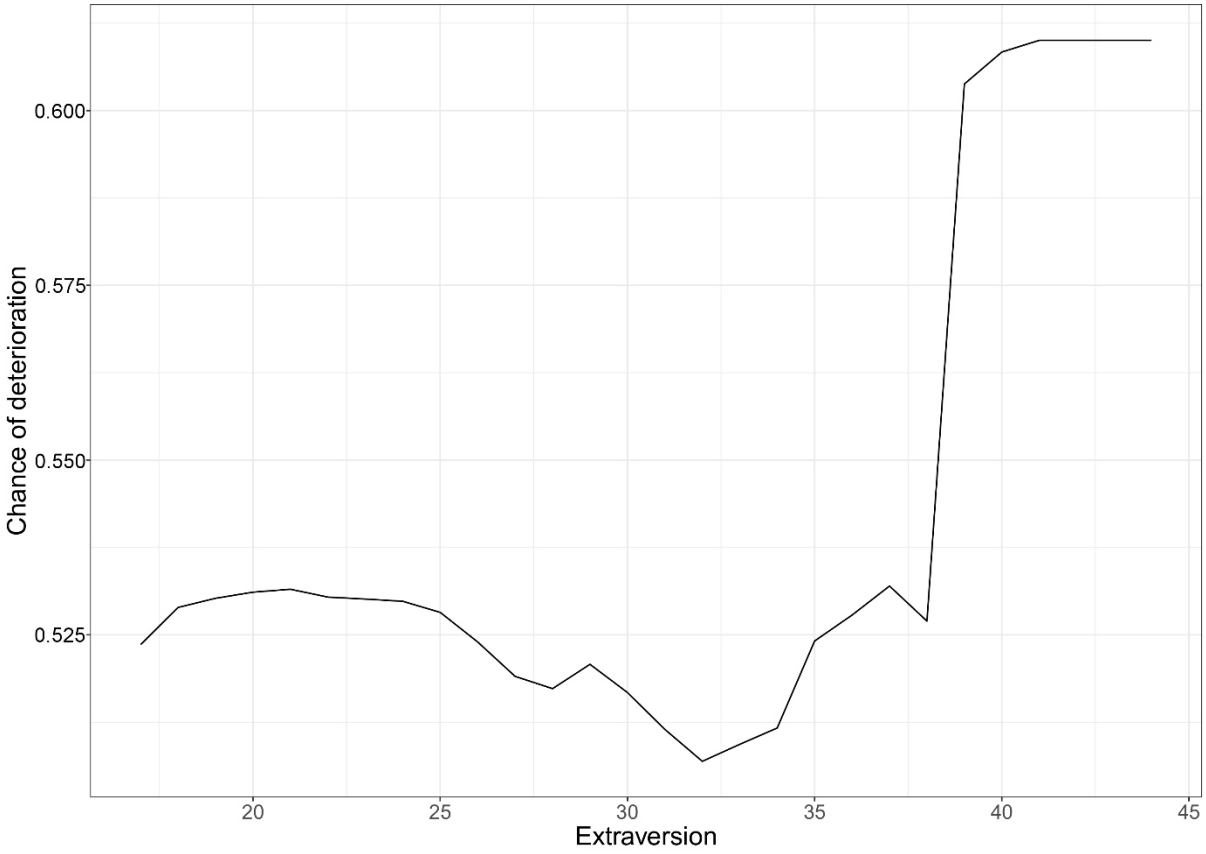
Partial dependency plot for satisfaction with partner relationship/sexuality



Note. Individual contribution of trainee satisfaction with their partner relationship/sexuality to the chance of deterioration. Higher values on the y-axis indicate a higher likelihood to deteriorate.

Figure 5

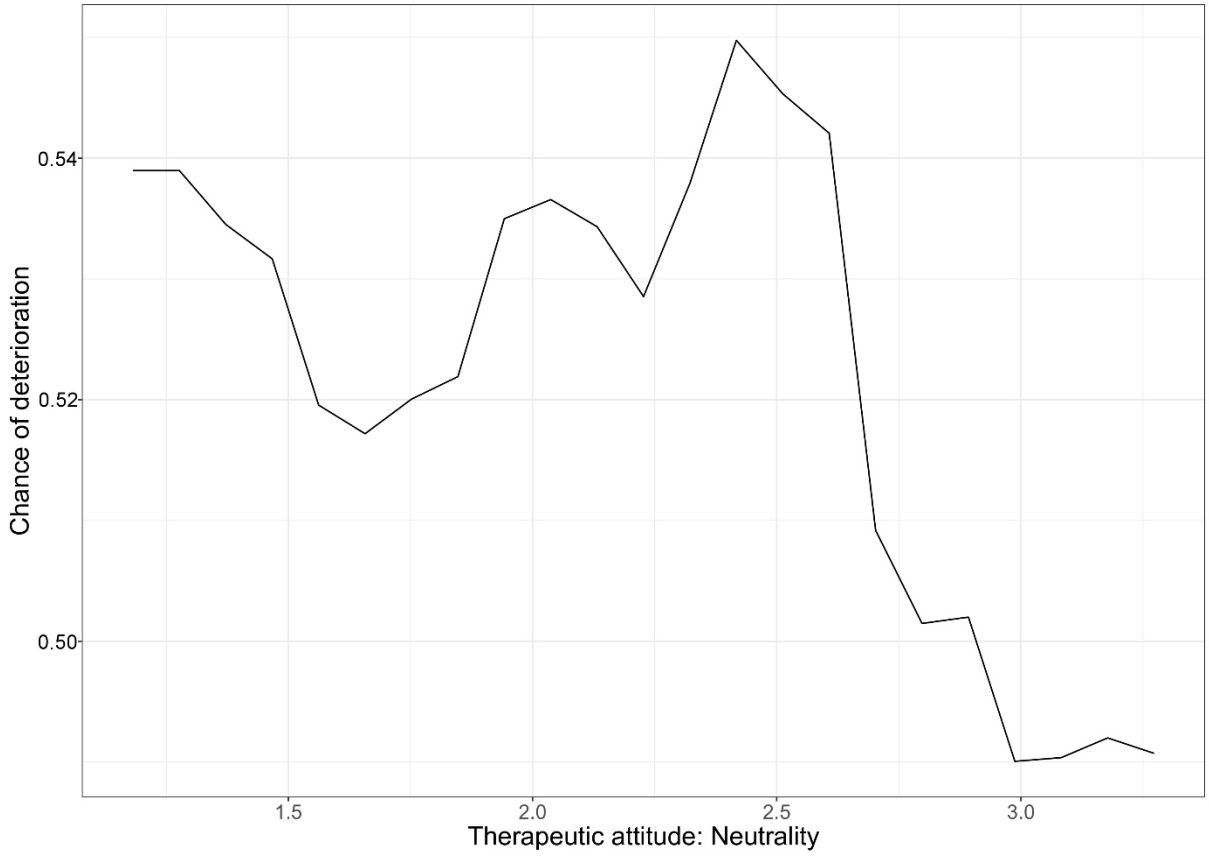
Partial dependency plot for extraversion



Note. Individual contribution of extraversion to the chance of deterioration. Higher values on the y-axis indicate a higher likelihood to deteriorate.

572 **Figure 6**

573 *Partial dependency plot for the therapeutic style "neutrality"*



574

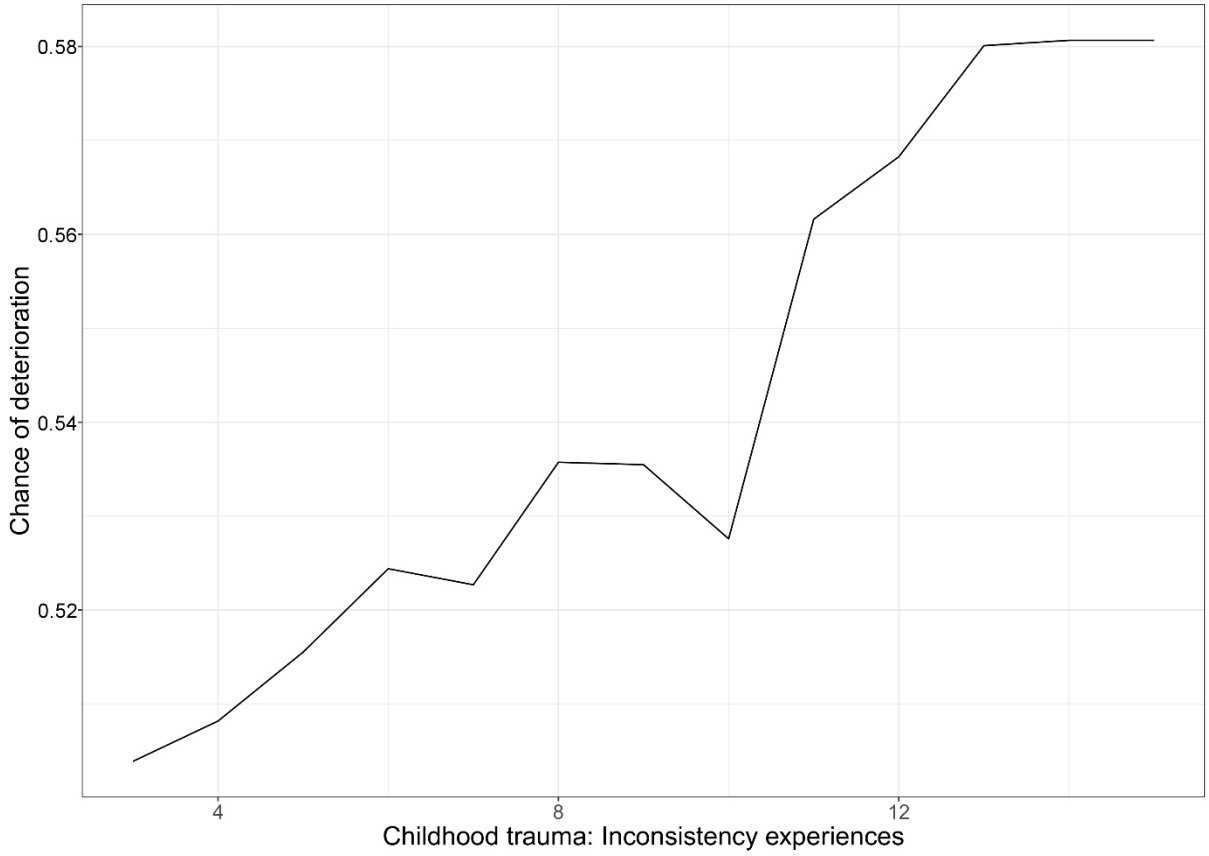
575

576 *Note.* Individual contribution of neutrality to the chance of deterioration. Neutrality represents one
577 scale of self-described therapeutic style as part of trainees' therapeutic attitude. Higher values on
578 the y-axis indicate a higher likelihood to deteriorate.

579

Figure 7

Partial dependency plot for inconsistency experiences in childhood



Note. Individual contribution of inconsistency experiences to the chance of deterioration. Inconsistency experiences describe self-reported unpredictable parenting behavior in childhood. Higher values on the y-axis indicate a higher likelihood to deteriorate.