Psychological Test and Assessment Modeling, Volume 64, 2022 (1), 3-21

Note: Machine Learning Modeling and Optimization Techniques in Psychological Assessment

David Goretzko[1] *& Markus Bühner*[2]

Abstract

Recently, machine learning modeling has made its way into psychological research. While it is used mostly in regression or classification contexts to optimize the prediction of certain variables, its principles and techniques also have arrived in psychometrics and psychological assessment. In this paper, we present machine learning and optimization concepts that can be used for different aspects of questionnaire development and test construction focusing on four central issues – item development and item selection, dimensionality assessment in latent variable modeling, improving the generalizability of factor models as well as the evaluation of measurement invariance or differential item functioning. By introducing different machine learning techniques and newly developed methods, we want to encourage researchers to try out these tools and upgrade their psychometrics toolboxes.

Keywords: Psychometrics; Regularization; Machine Learning; Latent Variable Modeling; Recursive Partitioning

[1] *Correspondence concerning this article should be addressed to:* David Goretzko, LMU Munich Department of Psychology, Leopoldstr. 13, 80802 Munich; email: david.goretzko@psy.lmu.de

[2] LMU Munich Department of Psychology

Introduction

Over the past years, machine learning (ML) has become more and more popular in psychological research. The powerful ML algorithms are used in various areas such as personality psychology (Stachl et al., 2020), clinical psychology (Dwyer et al., 2018), organizational psychology (Goretzko & Israel, 2022), educational psychology (Sinclair et al., 2021), and many more. Most of the time, ML models are applied in prediction settings, in which they promise to outperform classical (often linear) regression or classification models. When it comes to psychometrics, psychological assessment, and the measurement of latent variables, ML techniques and principles may also augment the standard toolbox and improve current practices. In this paper, we want to briefly introduce several methods and tools that have been developed recently to address common issues in psychological assessment and latent variable modeling - *A*) the item construction and item selection in questionnaire development and test construction, *B*) the dimensionality assessment in latent variable modeling (especially in exploratory factor analysis, EFA), *C*) the generalizability and interpretability of factor models as well as *D*) the detection of differential item functioning (DIF) and the investigation of measurement invariance. In the following, we want to discuss new developments for each of the four aspects (*A*) - *D*)) focusing on the potential of ML techniques or principles and closely related methods. In doing so, we want to encourage practitioners to try out these new methods and add them to their repertoire as they provide a new perspective on common challenges in psychological assessment.

The four aspects (*A*) - *D*)) cover different aspects of latent variable modeling that can be addressed with machine learning and optimization techniques. Psychological assessment can, of course, also benefit from directly applying ML models for prediction tasks (i.e., regression or classification tasks). For a broad overview of personality assessment with ML modeling, for example, and the use of new data sources such as digital footprints and mobile devices, we recommend the works of Bleidorn and Hopwood (2019), and Stachl et al. (2020, 2021). The assessment of emotions and cognitive appraisals that arguably play an important role in the experience and/or the emergence of emotions can also be improved by ML prediction models. The works of Meuleman and Scherer (2013), Israel and Schönbrodt (2020), or Zhang et al. (2020) illustrate how machine learning modeling can be used to predict emotional states, cognitive appraisals related to emotional experience or affective reactions on video stimuli (especially, when combined with physiological data). Besides emotion and personality assessment, new data sources in combination with ML modeling seem to be promising in all kinds of prediction contexts in psychological assessment. However, researchers have to avoid certain pitfalls that may lead to overly optimistic performance evaluation or yield severe misinterpretations (Orru et al., 2020; Yarkoni & Westfall, 2017). Since, review articles such as Orru et al., (2020), Stachl et al. (2020) or Pargent and Albert-von der Gönna (2019) provide detailed descriptions of predictive modeling approaches in psychological assessment, this paper focuses on the four aspects (*A*) - *D*)) that are more strongly related to developing measurement models for psychological constructs.

A) Item Construction and Item Selection

Item construction, as well as item selection, are usually tedious steps that require several runs of rephrasing, testing, and complex checks. Hence, developing psychological tests or questionnaires can take a lot of time and resources. Even though ML and optimization techniques will not make these steps redundant, they might help to reduce the effort of designing an item set and selecting the most suitable indicators.

Automated Item Generation

Over the past years, natural language processing, artificial text recognition, and text production methods have made huge progress (e.g., Jurafsky & Martin, 2019; Mitkov, 2014). Voice-operated assistants and automatic text translation are just two examples of how artificial entities have learned to "understand" speech and text. Trained on enormous corpora, modern natural language processing or natural language generation models are able to write journalistic texts (e.g., Dörr, 2016; Gatt & Krahmer, 2018), but also poetic texts that when pre-selected by a human were not distinguishable from poems written by human authors (Köbis & Mossink, 2021). These findings suggest that natural language generation models that are trained on an item set (e.g., a large pool of BIG-5 personality items) can be used to generate further items to develop a second item set for a parallel version of a scale, or simply to generate a large pool of items a researcher then can choose from. While for new or not well-established psychological constructs (where no large item pool has already been created) training a natural language generation models model does not seem to be too promising, it may be worth pursuing considering well-known constructs such as personality traits or clinical concepts. For psychological and knowledge tests where it is far easier to develop parallel versions and items with equal difficulties and inter-item correlations, the ML approach seems to be less appealing. In education settings, automatic item generation (AIG) based on "item models" that are able to variate certain features to create different versions of the same item (Gierl et al., 2012) seems to be the more auspicious approach. The same may be true for psychological tests - especially intelligence tests - where such AIG models are used to create figural analogy items (Blum & Holling, 2018), mazes (Blum & Holling, 2018) or Raven's Progressive Matrices (Wang & Su, 2015).

Natural language processing or natural language generation models, on the other hand, can become suitable AIG models in the context of psychological questionnaires as demonstrated by von Davier (2018). Since the long-short-term memory (LSTM) network trained by von Davier (2018) was not tailored to a specific trait (it was trained on a large corpus of various personality traits), Hommel et al. (2021) developed a transformer-based deep learning model that generates items for more specific (personality) traits such as risk-taking. Their work shows that natural language generation models have the potential to support researchers formulating items and generating a large item pool which may help to develop new scales but also simplifies the

development of parallel versions for existing scales that have comparable psychometric properties. Götz et al. (2021) pursue a very similar approach also relying on a transformer model to generate new items for psychological scales. Besides the possibility to use AIG models in the development of parallel forms of existing scales, the authors discuss their applicability for item generation in different languages and for regional varieties.

Short-Scale Development using Optimization Algorithms

While ML or deep learning models can be used for item generation, optimization techniques have been recently applied for item selection and short-scale development (Schroeders et al., 2016). With increasing computational power, fitting latent variable models becomes faster and easier every year. Hence, it is now possible to try out different item sets, manipulate items, etc., and refit a model with barely any time delay. However, the larger an item pool is and the more different objectives a researcher wants to achieve (e.g., an excellent model fit and high item reliabilities while having very diverse item difficulties at the same time), it becomes impossible to find the "optimal" set of variables "by hand". A greedy approach trying out all combinations of items becomes infeasible and the pattern of which items to add to a scale to better the model fit can be too complex for a researcher to read. Therefore, optimization algorithms such as the ant-colony algorithm have been used to find optimal item sets for shortscales (Olaru et al., 2015). The idea of this approach is inspired by the signaling system ants use when searching food resources. At first, the "ants" randomly explore different item sets, and those combinations that have desired psychometric properties get higher weights for subsequent runs - a so-called pheromone trail guides the following ants, so that they select the items that have shown good properties with a higher probability. Over time, suitable items get higher levels of pheromones and are selected more often than items that have undesirable properties. Due to its probabilistic nature, ant-colony optimization (ACO) is not likely to terminate in local minima but may yield different solutions for different starting points which is why it should probably be run several times. Olaru et al. (2019) suggest relying on cross-validation to assess the stability of the ACO solution as well as to avoid overfitting to the respective data set.

ACO was already used to develop a HEXACO short-scale (Olaru & Jankowsky, 2021) as well as measurement invariant short-scales (Jankowsky et al., 2020). Olaru et al. (2019) illustrate how ACO can be used to develop short-scales and to select items based on psychometric properties. In doing so, researchers are able to optimize the model fit of the respective measurement model, the predictive validity of the latent variable, or the item reliability. Usually, it is not meaningful to simply optimize one measure as the respective optimization procedure will most likely overfit and select an item set which cannot be considered representative for a specific psychological construct. Therefore, researchers should always consider different objectives for the ACO approach, while also not sacrificing the content validity. Focusing on a good

model fit and high reliability may result in an item set that is too narrow and does not cover all aspects of a psychological construct.

Content Validity

ML or deep learning modeling, as well as optimization techniques such as ACO, will probably assist researchers when developing questionnaires in future research. They can facilitate the troublesome process of designing items and selecting the most suitable ones from a large item pool. However, researchers applying these methods must have domain expertise and have to be aware of the limitations of the respective tools. Assuring content validity should always be of special interest when developing a new scale. Solely relying on automation and quantifiable psychometric properties that can be used for optimization procedures such as ACO is obviously not advisable. Nonetheless, adding these approaches to the "psychometrics toolbox" may help to improve psychological assessment.

B) Dimensionality Assessment in Latent Variable Modeling

In latent variable modeling, especially in EFA, assessing the dimensionality is a crucial and arguably the most far-reaching task a researcher faces. Determining the number of factors in the EFA has been a challenge for researchers for years and numerous so-called factor retention criteria have been developed to tackle this issue. Although some methods such as parallel analysis (Horn, 1965) or the minimum average partial test (Velicer, 1976) for principal component analysis have shown quite good results in a variety of simulation studies (e.g., Auerswald & Moshagen, 2019; Zwick & Velicer, 1986) and are therefore recommended by several authors (e.g., Fabrigar et al., 1999), research also indicates that there is no criterion among the established methods that reaches high accuracy under all data conditions (Goretzko et al., 2019). As an alternative to conventional criteria that are either based on simulated data such as parallel analysis or the similar comparison data approach (Ruscio & Roche, 2012) or reference eigenvalues that are calculated considering the sample size and other data characteristics such as the empirical Kaiser criterion (Braeken & Van Assen, 2017), ML-based methods for dimensionality assessment have been developed - the so-called factor forest approach by Goretzko and Bühner (2020, 2022a) and exploratory graph analysis by Golino and Epskamp (2017). In the following, we want to briefly introduce these two new approaches to factor retention in EFA. Lately, ML-based methods for parameter estimation in factor analytic models have also been developed (e.g., in exploratory item factor analysis, Urban & Bauer, 2021). As these methods rely on variations of common factor retention criteria (e.g., a tailored version of the Scree-test, Urban & Bauer, 2021) to determine the number of latent factors, we do not discuss these approaches in detail.

The Factor Forest

The factor forest is a method that combines extensive data simulation with ML modeling to create a prediction model with a supervised learning approach (Goretzko & Bühner, 2020). The basic idea is to build a model on simulated data that (fully) covers the relevant conditions of an application context with regard to sample sizes, numbers of indicators, numbers of latent variables, communalities, and loading patterns. For simulated data, the true dimensionality (i.e., the true number of factors) is known, so the ML model can "learn" the relationship between several data features (i.e., characteristics of the empirical data that can be retrieved or calculated for every data set) and the number of latent factors. Goretzko and Bühner (2020) simulated approximately 500,000 data sets varying the sample size, the number of latent and manifest variables, the loading pattern, and the inter-factor correlations and calculated 184 features for each data set.

They used the empirical eigenvalues of the correlation matrix, eigenvalues of a reduced correlation matrix (based on the common factor model), matrix norms of the correlation matrix, inequality measures (such as the Gini-coefficient), and other summary statistics based on the manifest correlations as well as general data characteristics (e.g., the sample size, the number of indicators) as features and trained a gradient boosting algorithm (the *XGBoost*, Chen et al., 2018) to predict the number of factors given the respective features.

Their trained model reached an out-of-sample accuracy above 99% and outperformed common criteria such as parallel analysis or the empirical Kaiser criterion (Goretzko & Bühner, 2020). This superiority of the pre-trained machine learning model to common factor retention criteria was not only found for multivariate normal data (i.e., data conditions comparable to the training context) but also for ordinal data (Goretzko & Bühner, 2022a). In addition, the authors showed that the factor forest (i.e., the trained

XGBoost model) not only performs well in simulation studies but also reached higher replicability rates in a study with empirical data (Goretzko & Bühner, 2022b). Hence, the factor forest approach seems to be promising when it comes to determining the number of factors for an EFA and can be seen as an indication that ML modeling may help to improve latent variable modeling.

Exploratory Graph Analysis

Exploratory graph analysis is another promising alternative for dimensionality assessment in factor analyses, even though it is based on network modeling where no latent variables are explicitly modeled (for a general introduction to the research area combining network modeling and psychometrics, see, Epskamp et al., 2018 as well as Epskamp, 2021). Contrary to EFA and traditional latent variable modeling in psychology, exploratory graph analysis is based on a network model called the Gaussian graphical model (e.g., Epskamp et al, 2018). Within such a network, each node represents a manifest variable (e.g., a questionnaire item) and the (standardized) edges

(connections between two nodes) reflect the partial correlation between the respective variables. In exploratory graph analysis, the Gaussian graphical model is not estimated as such (based on the inverted variance-covariance matrix), but the Likelihood is penalized with a regularization term (for further readings on the graphical LASSO, see Friedman et al., 2007) to stabilize the parameter estimates and improve generalizability (Golino & Epskamp, 2017). This approach is similar to regularized EFA and regularized structural equation modeling (SEM) which we describe in the next section ("C: The Generalizability and Interpretability of Factor Models"). This regularization shrinks the partial correlations towards zero and fosters a sparser network (some edge coefficients become zero) which promises to be more replicable across samples (Golino & Epskamp, 2017). Within this sparser network, some clusters of variables may form dense subgraphs (also called "communities" in network analyses). Exploratory graph analysis relies on a walktrap algorithm (Pons & Latapy, 2006) to determine the number of clusters (or subgraphs) that are formed by subsets of the manifest variables. Christensen et al. (2020) argue that each of these clusters can be seen as a representation of an underlying factor, i.e., that the approach can be used to determine the dimensionality of factor models as well. Golino et al. (2020) show that its performance is comparable to standard procedures such as parallel analysis and the minimum average partial test and even outperforms them in conditions with high inter-factor correlations. Hence, exploratory graph analysis can be an alternative tool to determine the dimensionality for both questionnaire (Christensen et al., 2020) and cognitive test data (Golino & Demetriou, 2017).

Applicability of the new Approaches

Both the factor forest and exploratory graph analysis can be seen as a valuable extension of the factor retention toolbox as they are able to accurately determine the number of factors. While the factor forest promises very high accuracy (arguably the highest accuracy by a stand-alone method) when the application context is covered appropriately for the model training, exploratory graph analysis additionally provides information about the inter-relations of the indicator variables which can be used for visualizations (e.g., by plotting the sparse network structure with the different communities). The latter can be helpful for more detailed interpretations and may be used for item diagnostics.

C) The Generalizability and Interpretability of Factor Models

The replicability and generalizability of factor solutions is of great interest in psychological assessment. It has become common practice that EFA results are validated on new data using confirmatory factor analysis (CFA; Fabrigar et al., 1999; Goretzko et al., 2019) or in a second semi-exploratory step using the exploratory structural equation modeling framework (Asparouhov & Muthén, 2009). The latter has been developed as CFA models (especially CFA models implying simple structure) often do not fit the data properly (Hopwood & Donnellan, 2010). However, replication attempts with exploratory structural equation modeling also reveal that factor structures found in one sample using classical EFA, cannot be transferred to another sample. In other words, EFA results - especially when rather small samples were used - frequently lack (exact) replicability. One reason for this issue may be the signal-to-noise ratio in questionnaire measures (e.g., Gnambs, 2015) in combination with rather small sample sizes. One possible solution could be regularized latent variable models that trade variance against small biases in parameter estimation (for more on this bias-variance trade-off, see, for example, Yarkoni & Westfall, 2017).

Regularized Exploratory Factor Analysis

In regularized EFA, the log-likelihood that is maximized with respect to both loading parameters Λ and unique variances Ψ^2 in common maximum likelihood estimation (ML-EFA) is completed by a penalty term ($P(|\lambda_{ij}|)$) that penalizes non-zero parameter estimates (here non-zero factor loadings):

$$l_{pen}(\Lambda, \Psi^2) = -\frac{N}{2}[p\log(2\pi) + \log|\Lambda\Lambda^T + \Psi^2| + tr(|\Lambda\Lambda^T + \Psi^2|^{-1}S)] - N\sum_{i=1}^{p}\sum_{j=1}^{k}\gamma\, P(|\lambda_{ij}|)$$

where p is the number of manifest variables, k the number of latent variables, N the sample size, S the sample variance-covariance matrix, and γ a regularization parameter that controls the amount of shrinkage (the higher it becomes, the more strongly the parameter estimates will be reduced).

One can see that the penalty term lowers the log-likelihood which means it somewhat makes it "more difficult" to maximize the term (to be more precise: it ensures that a solution with smaller or fewer non-zero loading parameters maximizes the penalized log-likelihood). Hirose and Yamamoto (2014) use the so-called MC+ penalty to obtain a factor solution with loading patterns that are as sparse as possible (i.e., a solution with very few cross-loadings and as many zeros as possible in $\widehat{\Lambda}$). While this penalty might be a good choice when very sparse loading matrices can be expected (Hirose & Yamamoto, 2015), which might be the case when numerous heterogeneous manifest variables are used without the purpose to develop a measurement model for a specific latent construct (e.g., in panel surveys that collect data on various topics), other penalties seem to be more appropriate for psychological data and psychological assessment in particular. Scharf and Nestler (2019) evaluated LASSO, Ridge, and

ElasticNet penalties in data conditions that are more common in psychological research settings and found the ElasticNet penalty which is a combination of LASSO and Ridge regularization to perform best (for more information on the LASSO penalty as well as the idea of the ElasticNet, see Tibshirani, 2011 and Zou and Hastie, 2005).

Besides decreasing estimation variance and thereby potentially increasing the replicability of the factor structure, regularized EFA can be used to foster the interpretability of the loading patterns. While common EFA usually requires a two-step approach extracting the factors and then rotating the initial factor solution to obtain interpretable patterns, regularized EFA may be a promising alternative to factor rotation (Goretzko et al., 2019; Scharf & Nestler, 2019). Choosing a rotation method is always an ambiguous and very challenging task for a researcher as there is no data-driven way to determine a "correct rotation" (Browne, 2001; Schmitt & Sass, 2011). In fact, rotational indeterminacy (i.e., the factor solution is only determined exactly up to a permissible rotation, and an infinite set of factor loadings and inter-factor correlations fit the data equally well) can only be "resolved" by theoretical considerations (Browne, 2001). Adding a penalty term to the likelihood function removes this problem, as regularized EFA results are unique up to the order of factors and sign switches (Scharf & Nestler, 2019). Thus, researchers might consider regularized EFA as an alternative to common EFA that does not require them to select an appropriate rotation method for better interpretability,while potentially improving the replicability and generalizability of the results.

Regularized Structural Equation Modeling

The concept of regularization via penalized maximum likelihood estimation can also be transferred to the SEM framework (Jacobucci et al., 2016). The log-likelihood or a related objective function of the maximum likelihood estimation in SEM can be completed by a penalty term analogously to the regularized EFA approach (see above). Hence, for questionnaire development or similar purposes, CFA models can also be fitted with penalized maximum likelihood to obtain more generalizable results (Li et al., 2021). As regularization results can be unstable - especially when sample sizes are small - stability selection methods have been developed to tackle this issue (Li & Jacobucci, 2021).

Within the regularized SEM framework, different kinds of parameters can be regularized. Depending on theoretical considerations and previous study results, it is possible to apply the penalty to all model parameters, but also to the regression coefficients, inter-factor correlations or loadings separately (Jacobucci et al., 2021; Li et al., 2021). In doing so, researchers are able to only penalize the regression coefficients that describe relations between a latent variable and several external manifest variables to get an idea of which covariates or criteria are related with a specific construct, while not regularizing the measurement model which is already theoretically well-founded, for example.

Just like regularized EFA, regularized SEM is also useful with many manifest variables (Deng et al., 2018) and in small sample size scenarios (e.g., Jacobucci et al., 2019).

While the LASSO penalty can be unstable with small samples, though (Li & Jacobucci, 2021), its variable selection properties come into play in large-scale surveys with numerous variables that are either integrated into a measurement model or are assumed to be related to different latent variables. In panels with comparably high sample sizes, regularized SEM promises to carve out patterns that should be more robust to sampling errors than those provided by common SEM.

Regularization in Latent Variable Modeling

First simulation studies (e.g., Scharf & Nestler, 2019) show auspicious results for regularization in latent variable modeling. As described above, both regularized EFA and regularized SEM can add valuable aspects to the psychometric toolbox, as they promise to increase the replicability and generalizability of factor solutions, while also providing an alternative to the two-step EFA approach that requires researchers to select a rotation method.

D) Differential Item Functioning and Measurement Invariance

Measurement invariance and differential item function (DIF) are always a concern when developing psychological scales and constructing tests. While there are established methods for both item response theory (IRT) and the SEM framework (e.g., Meade & Lautenschlager, 2004; Schoot et al., 2012), these approaches require researchers to define subgroups for which measurement invariance is tested or to select variables that are considered when modeling DIF. Hence, detecting DIF and non-invariance can be challenging, especially when researchers want to address it right from the beginning of the test construction process (i.e., considering measurement invariance during item selection). ML tools, first and foremost model-based recursive partitioning, can help to tackle DIF and measurement invariance during questionnaire development. In the following, we want to briefly talk about IRT model trees that apply recursive partitioning to IRT models which can be used to detect DIF and related subgroups for which the model structure differs. We also present SEM trees that combine recursive partitioning and SEM which can be used to find subgroups with differing measurement models (or structural models).

IRT Model Trees and Regularized IRT Models

IRT model trees (not to be confused with IRTrees that are designed to model multiple response processes in categorical data, see, for example, Plieninger, 2021) are able to detect DIF in IRT models exploring several covariates and interactions or

combinations of covariates effectively without requiring the researcher to define subgroups in advance. This data-driven way of assessing DIF can be applied automatically to models for dichotomous variables with Rasch model trees (Strobl et al., 2015) as well as polytomous data with partial credit trees (El-Komboz et al., 2014; Komboz et al., 2018). In both cases, the respective tree is built in four steps. First, the model parameters (e.g., threshold and item parameters in the partial credit model) are estimated for the full sample. Then, the stability of the item and/or threshold parameters is assessed given all covariates (e.g., gender or age that are usually considered in DIF analyses, but also other variables that are collected for this sample such as personality measures can be used here). In a third step, if instabilities are deemed significant (for further information about the internal significance tests, see Komboz et al., 2018; Strobl et al., 2015; Zeileis et al., 2008), the ideal split point is determined for the covariate that promises the greatest improvement in model fit and the sample is divided into two subsamples (binary splitting). These steps are then repeated until no significant improvements (i.e., no further splits) can be found. This way, a tree structure is grown that divides the sample into several groups, for which the parameters of a Rasch or partial credit model are estimated separately. In case no DIF is present, no splitting should be done and the respective model is fitted to the whole data set, but if there are subgroups that substantially differ with regard to item comprehension, for example, the Rasch model or partial credit trees promise to find the covariates or interactions of covariates and the best cut-points in numeric DIF variables that best separate these subgroups (e.g., Strobl et al., 2015).

A different approach to detecting DIF in IRT models has been suggested by Tutz and Schauberger (2015). Their idea was to replace the item parameters of the model with a linear term that includes covariates that potentially explain DIF (typically, variables such as age, gender, etc.). By penalizing the log-likelihood (see also the discussion of regularization in the paragraphs above), the resulting model with way more parameters than the initial Rasch model can still be estimated. Using a LASSO penalty with its variable selection property (e.g., Tibshirani, 2011), all regression coefficients that belong to covariates which cannot explain DIF are shrunk to zero. Schauberger and Mair (2020) extended the approach to the ordinal (generalized) partial credit model, while Belzak and Bauer (2020) evaluate its potential to identify DIF in the 2-PL model. A slightly different, yet conceptually similar approach was proposed by Magis et al. (2015). Contrary to Tutz and Schauberger (2015), they use the item sum-scores as a proxy for the latent variable and model DIF with a logistic regression with LASSO penalty. An advantage of these regularization approaches is that DIF is directly modeled and that a functional relationship between DIF covariates and the actual item parameters (or item answers in case of Magis et al., 2015) can be established. The IRT model trees, on the other hand, explicitly reveal subgroups that differ with regard to the model parameters and do not need to set the type of relationship (linear or non-linear) between DIF covariates and item or threshold parameters.

SEM Trees

Model-based recursive partitioning can also be used in the SEM framework to build so-called SEM trees (Brandmaier et al., 2013). While measurement invariance is usually tested with multi-group CFA (French & Finch, 2008), SEM trees promise to detect subgroups defined by one or more covariates and their interactions for which the measurement models differ (Finch, 2017). Although SEM trees are intended to predominantly assess subgroup differences in the structural model (i.e., to find subgroups for which the relations among latent variables differ depending on covariates such as age, gender, or personality), it can also be used as a more exploratory approach to measurement invariance testing by focusing on the measurement models of latent variables. The general idea of SEM trees is quite similar to the IRT model trees described above.

First, the SEM model is fitted to the complete data set (the SEM could be a simple factor model to perform a CFA if the researcher wants to explore measurement [non]invariance). In a second step, instabilities in the parameter estimates with regard to all covariates are assessed, and then, the data set is split given the covariate and the respective cut-point that yield the highest improvement in model fit. This procedure is repeated recursively until no more significant improvements can be made (Brandmaier et al., 2013).

Both Rasch model trees (Strobl et al., 2015) and SEM trees (Brandmaier et al., 2013), as well as other approaches relying on model-based recursive partitioning, face the problem of multiple testing when determining which splits lead to a significant improvement in model fit and when to stop growing the tree. Since the common Bonferroni correction method is rather conservative and adversely affects the statistical power to find relevant subgroups, Brandmaier et al. (2013) prefer cross-validating SEM trees when determining split points and addressing the problem of multiple testing. More generally speaking, cross-validation may also help to find a stable solution that replicates and generalizes to other samples in different settings. Single trees are often regarded as somewhat unstable (James et al., 2013) which is why Brandmaier et al. (2016) also developed an ensemble method combining several trees - the so-called SEM forests. When exploring differences in the structural model of an SEM, high stability and replicability may be of uppermost interest, so that SEM forests are arguably more powerful than simple SEM trees. However, when it comes to finding subgroups with different measurement models - which can be seen as the first step in a measurement invariance analysis as well as an inherently exploratory approach, SEM trees are way easier to interpret and define these subgroups by providing clear cut-points for selected covariates. SEM forests, on the contrary, are based on averaged decision trees and are therefore not accessible for researchers. Accordingly, SEM trees may be preferable for measurement invariance analyses.

EFA-Trees

The issue of measurement invariance is often not considered during the first steps of questionnaire development, because there are no statistical tools for investigating measurement invariance at an early stage of this process. When selecting variables and redefining a latent concept using EFA, measurement invariance is not addressed. The SEM tree or model-based recursive partitioning framework can be used to change that, though. Instead of a CFA model, the common EFA model can be integrated with the model-based recursive partitioning framework to explore measurement invariance when selecting items and doing the first solely exploratory analyses during test construction. The procedure of such an EFA tree would be comparable to the SEM tree. First, the EFA model is estimated for the full sample. Then, instabilities in the parameter estimates (e.g., the factor loadings) are assessed, and in case of significant instability, the best split variable and the ideal cut-point are used to find subgroups. This process is repeated as long as significant improvements in model fit are detected.

Recursive Partitioning for Exploratory Analyses

Using model-based recursive partitioning, trees can be built that detect and illustrate DIF or non-invariance of measurement models. The approach has been adopted in both IRT and classical test theory making it useful for dichotomous, polytomous as well as continuous variables. While common confirmatory approaches such as multigroup CFA are still necessary to test for measurement invariance or DIF, these tree-based methods can be really powerful in finding subgroups and developing hypotheses which groups to test in a subsequent confirmatory analysis on new data.

Conclusion

Although ML modeling and related techniques will not entirely change psychometrics by solving all problems with common methods, they promise to enrich our toolbox when designing psychological scales and modeling latent concepts. This paper is not intended as a call to immediately implement all these methods and completely rely on ML-based psychometrics but rather to invite researchers to explore new directions for familiar challenges. We are convinced that the presented ML concepts and modern methods are meaningful ways to tackle these challenges and that the psychological assessment research will benefit from applying them. Nonetheless, it has to be clearly stated that, no matter how promising these methods may be, psychological domain knowledge will still be needed - especially considering the assessment of (content) validity and the relevance of specific measures.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*(4), 468–491. https:// doi.org/10.1037/met0000200

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi .org/10.1037/met0000253

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

Blum, D., & Holling, H. (2018). Automatic generation of figural analogies with the IMak package. *Frontiers in Psychology*, *9*, 1286.

Braeken, J., & Van Assen, M. A. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*(3), 450–466. https://doi.org/10.1037/met0000074

Brandmaier, A. M., Oertzen, T. von, McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https:// doi.org/10.1037/a0030001

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theoryguided exploration with structural equation model forests. *Psychological Methods*, *21*(4), 566–582. https://doi.org/10.1037/met0000090

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2018). Xgboost: Extreme gradient boosting. R package version 0.6. 4.1.

Christensen, A. P., Golino, H., & Silvia, P. J. (2020). A psychometric network perspective on the validity and validation of personality trait questionnaires. *European Journal of Personality*, *34*(6), 1095–1108. https://doi.org/10.1002/per .2265

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, 580. https://doi.org/10.3389/fpsyg.2018.00580

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, *4*(6), 700–722. https://doi.org/10.1080/21670811.2015.1096748

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816 -045037

El-Komboz, A., Zeileis, A., & Strobl, C. (2014). *Detecting differential item and step functioning with rating scale and partial credit trees*. University of Munich, Department of Statistics. https://doi.org/10.5282/ubm/epub.17984

Epskamp, S. (2021). *Psychonetrics: Structural equation modeling and confirmatory network analysis*. Retrieved from https://CRAN.R-project.org/package=psychonetrics

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195– 212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018 .1454823

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3 .272

Finch, W. H. (2017). Structural equation modelling trees for invariance assessment. *International Journal of Quantitative Research in Education*, *4*(1-2), 72–93. https://doi.org/10.1504/IJQRE.2017.086508

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 96–113. https://doi.org/10.1080/10705510701758349

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*(8), 757–765. https:// doi.org/10.1111/j.1365-2923.2012.04289.x

Gnambs, T. (2015). Facets of measurement error for scores of the big five: Three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89. https://doi.org/10.1016/j.paid.2014.08.019

Götz, F. M., Maertens, R., & Linden, S. (2021). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *PsyArXiv*. https://doi.org/10.31234/osf.io/m6s28

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), e0174035.

Golino, H. F., & Demetriou, A. (2017). Estimating the dimensionality of intelligence like data using exploratory graph analysis. *Intelligence*, *62*, 54–70. https:// doi.org/10.1016/j.intell.2017.02.007

Golino, H. F., Shi, D., Christensen, A. P., Garrido, L. E., Nieto, M. D., Sadana, R., ... Martinez-Molina, A. (2020). Investigating the performance of exploratory graph analysis and traditional techniques to identify the number of latent factors: A simulation and tutorial. *Psychological Methods*, *25*(3), 292–320. https://doi.org/10.1037/met0000255

Goretzko, D., & Bühner, M. (2020). One model to rule them all? Using machine learning algorithms to determine the number of factors in exploratory factor analysis. *Psychological Methods*, *25*(6), 776–786. https://doi.org/10 .1037/met0000262

Goretzko, D., & Bühner, M. (2022a). Factor retention using machine learning with ordinal data. *Applied Psychological Measurement*.

Goretzko, D., & Bühner, M. (2022b). Robustness of factor solutions in exploratory factor analysis. *Behaviormetrika*, *49*, 131–148. https://doi.org/10.1007/ s41237-021-00152-w

Goretzko, D., & Israel, L. S. F. (2022). Pitfalls of machine learning-based personnel selection: Fairness, transparency, and data quality. *Journal of Personnel Psychology*, *21*(1), 37–47. https://doi.org/10.1027/1866-5888/a000287

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*. https://doi.org/10.1007/s12144-019 -00300-2

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120– 132. https://doi.org/10.1016/j.csda.2014.05.011

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*(5), 863–875. https://doi.org/10.1007/s11222-014-9458-0

Hommel, B., Wollang, F.-J., Kotova, V., Zacher, H., & Schmukle, S. (2021). Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*. https://doi.org/10.1007/s11336-021-09823-9.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Israel, L. S. F., & Schönbrodt, F. D. (2020). Predicting affective appraisals from facial expressions and physiology using machine learning. *Behavior Research Methods*, *53*, 574–592. https://doi.org/10.3758/s13428-020-01435-y

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science*, *2*(1), 55–76. https://doi.org/10.1177/2515245919826527

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., Serang, S., Kievit, R. A., Scharf, F., ... Ye, A. (2021). *Regsem: Regularized structural equation modeling*. Retrieved from https://CRAN.R-project.org/package=regsem

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York, NY: Springer.

Jankowsky, K., Olaru, G., & Schroeders, U. (2020). Compiling measurement invariant short scales in cross–cultural personality assessment using ant colony optimization. *European Journal of Personality*, *34*(3), 470–485. https:// doi.org/10.1002/per.2260

Jurafsky, D., & Martin, J. H. (2019). *Speech and language processing (3rd draft ed.)*. Stanford University.

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166. https://doi.org/10.1177/0013164416664394

Köbis, N., & Mossink, L. D. (2021). Artificial intelligence versus maya angelou: Experimental evidence that people cannot differentiate AI-generated from humanwritten poetry. *Computers in Human Behavior*, *114*, 106553. https:// doi.org/10.1016/j.chb.2020.106553

Li, X., & Jacobucci, R. (2021). Regularized structural equation modeling with stability selection. *Psychological Methods*. https://doi.org/10.1037/met0000389

Li, X., Jacobucci, R., & Ammerman, B. A. (2021). Tutorial on the use of the regsem package in r. *Psych*, *3*(4), 579–592. https://doi.org/10.3390/psych3040038

Magis, D., Tuerlinckx, F., & de Boeck, P. (2015). A detection of differential item functioning using the Lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135. https://doi.org/10.3102/1076998614559747

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meuleman, B., & Scherer, K. R. (2013). Nonlinear appraisal modeling: An application of machine learning to the study of emotion production. *IEEE Transactions on Affective Computing*, *4*(4), 398–411. https://doi.org/10.1177/ 1094428104268027

Mitkov, R. (2014). *The Oxford Handbook of Computational Linguistics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199573691.001 .0001

Olaru, G., & Jankowsky, K. (2021). The HEX-ACO-18: Developing an age-invariant HEX-ACO short scale using ant colony optimization. *Journal of Personality Assessment*, *0*(0), 1–12. https://doi.org/10.1080/00223891.2021 .1934480

Olaru, G., Schroeders, U., Hartung, J., & Wilhelm, O. (2019). Ant colony optimization and local weighted structural equation modeling. A tutorial on novel item and person sampling procedures for personality research. *European Journal of Personality*, *33*(3), 400–419. https://doi.org/10.1002/per.2195

Olaru, G., Witthöft, M., & Wilhelm, O. (2015). Methods matter: Testing competing models for designing short-scale big-five assessments. *Journal of Research in Personality*, *59*, 56–68. https://doi.org/10.1016/j.jrp.2015.09 .001

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02970

Pargent, F. & Albert-von der Gönna, J. (2019). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie*, *226*(4), 246–258. https://doi.org/ 10.1027/2151-2604/a000343

Plieninger, H. (2021). Developing and applying IR-tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods*, *24*(3),654–670. https://doi.org/10.1177/1094428120911096

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *24*(3), 191–218. https://doi.org/10.1.1.262.4960

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*(2), 282–292. https://doi.org/10.1037/ a0025697

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576–590. https://doi.org/10.1080/10705511 .2018.1558060

Schauberger, G., & Mair, P. A. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behavior Research Methods*, *52*, 279–294. https://doi.org/10.3758/s13428-019 -01224-2

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71*(1), 95–113. https://doi.org/10.1177/0013164410387348

Schoot, R. van de, Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Schroeders, U., Wilhelm, O., & Olaru, G. (2016). Meta-heuristics in short scale construction: Ant colony optimization and genetic algorithm. *PloS One*, *11*(11), e0167110. https://doi.org/10.1371/journal.pone.0167110

Sinclair, J., Jang, E. E., & Rudzicz, F. (2021). Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing. *Journal of Educational Psychology*. https://doi.org/10 .1037/edu0000658

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., ... Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*, *34*(5), 613–631. https://doi.org/10.1002/per.2257

Stachl, C., Boyd, R. L., Horstmann, K. T., Khambatta, P., Matz, S., & Harari, G. M. (2021). Computational personality assessment. *Personality Science*, *2*, 1–22. https://doi.org/10.5964/ps.6115

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the rasch model. *Psychometrika*, *80*(2), 289– 316.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective.*Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771 .x

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43.

Urban, C. J., & Bauer, D. J. (2015). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, *86*(1), 1–29.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321–327. https://doi.org/10.1007/BF02293557

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika*, *83*(4), 847–857.

Wang, K., & Su, Z. (2015). Automatic generation of raven's progressive matrices. In *Twenty-fourth international joint conference on artificial intelligence*.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multimodal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103–126.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432