Comparing Model Comparison Methods

Holger Schultheis (schulth@informatik.uni-bremen.de)
Cognitive Systems, University of Bremen, Enrique-Schmidt-Str. 5, 28359 Bremen, Germany

Ankit Singhaniya
Computer Science and Engineering, NIT Nagpur, Nagpur 440010, India

Devendra Singh Chaplot
Computer Science and Engineering, IIT Bombay, Mumbai 400076, India

Abstract

Comparison of the ability of different computational cognitive models to simulate empirical data should ideally take into account the complexity of the compared models. Although several comparison methods are available that are meant to achieve this, little information on the differential strengths and weaknesses of these methods is available. In this contribution we present the results of a systematic comparison of 5 model comparison methods. Employing model recovery simulations, the methods are examined with respect to their ability to identify the model that actually generated the data across 3 pairs of models and a number of comparison situations. The simulations reveal several interesting aspects of the considered methods such as, for instance, the fact that in certain situations methods perform worse than model comparison neglecting model complexity. Based on the identified method characteristics, we derive a preliminary recommendation on when to use which of the 5 methods.

Keywords: computational cognitive models, model comparison, model mimicry, model generalization

When computationally modeling cognition, often several different models are available or conceivable as explanations for the cognitive ability in question. In such a situation, the aim is to select the best of these candidate models according to a set of criteria. Among others (e.g., falsifiability or interpretability) the extent to which the different models are able to simulate observed human behavior is usually considered a key criterion for selecting from the candidate models.

A naïve approach to gauge the models' ability to simulate the existing observations is to fit each model to the available data and choose the model that provides the tightest fit as indicated, for instance, by the models' Root Mean Squared Error (RMSE). Such an approach is problematic, because it does not take into account the complexity of the compared models. As a result, there is a tendency for overfitting and for selecting more complex models even if simpler models provide the better explanation of the considered cognitive ability (Pitt & Myung, 2002).

Several methods taking into account model complexity have been proposed to avoid the pitfalls of the naïve approach (see Shiffrin, Lee, Kim, & Wagenmakers, 2008, for an overview). However, common use of such more sophisticated model comparison methods is partly hampered by the fact that many properties of the different methods are insufficiently investigated. Only very few studies (e.g., Cohen, Sanborn, & Shiffrin, 2008) have systematically examined different comparison methods with respect to their differential advantages and disadvantages. Consequently, when faced with a situation that requires comparing models regarding their ability for simulating human behavior, modelers are often faced with the problem that it is unclear which model comparison methods could reasonably and should ideally be employed in a given situation.

In this contribution we present the results of a systematic comparison of 5 model comparison methods. The methods are examined with respect to their ability to select the model that actually generated the data across 3 pairs of models and a number of contextual variations (e.g., tightness of fits, amount of noise in the data). The obtained results highlight important properties of the different comparison methods. Together with the fact that all 5 considered methods are general in the sense that they place no restrictions on the type of models that can be compared, these results are, we believe, conducive to increasing the frequency with which more sophisticated comparison methods instead of the naïve approach will be employed for model evaluation and comparison.

The remainder of this article is structured as follows. First, we list and briefly describe all considered methods. Second, the employed models, contextual variations, and procedural details of the method comparison are described. Subsequently, comparison results are presented and discussed before we conclude our considerations and highlight topics for future work.

Methods

The 5 methods we compared are the bootstrap, the bootstrap with standard error (SE) and confidence interval (CI), the data-uninformed parametric bootstrap cross-fitting method, henceforth called cross-fitting method (CM), the simple hold-out, and the prediction error difference method (PED). Each of these was applied to 3 pairs of models and will be described in turn below.

Bootstrap

Given a set of n observations, the bootstrap method of model comparison proceeds as follows (see Efron & Tibshirani, 1993, for an overview of bootstrapping procedures). First, an arbitrary but fixed number B of bootstrap samples is generated. A bootstrap sample is a set of n data points randomly drawn with replacement from the n original observations. Due to sampling with replacement, most bootstrap samples will contain only a subset of all original observa-

tion (but some of these more than once). Second, each of the to-be-compared models is fitted to each bootstrap sample. Third, for each bootstrap sample, the fitted models are used to predict those data points that were not in the bootstrap sample and the deviation of the predictions from the original data points is measured (e.g., by the mean squared error). Fourth, the measures of deviation are combined for each model across all bootstrap samples to obtain an overall measure for the prediction error ($\bar{Err}$) of each model. The model that has the lowest $\bar{Err}$ is assumed to be the best approximation to the process that actually generated the n original data points.

Due to the randomness in generating the bootstrap samples as well as the noise that is likely included in the original observations, $\bar{Err}$ only constitutes an estimate of the models' true prediction error. Accordingly, the model showing the lowest $\bar{Err}$ may do so because of chance and not because it is the best model. Knowing the variability, that is, the SE, of the error estimates can potentially help alleviating this problem. Given the standard error, CIs on the true prediction error can be derived. If the CIs of the models' error estimates do not overlap, one may conclude with more confidence—depending on the confidence level employed to construct the intervals—that the model with the lower $\bar{Err}$ in fact provides the better approximation to the process that generated the n original data points.

In our simulations we assess both the bootstrap considering the SE and the bootstrap not considering the SE for deciding which of the two models is more appropriate. We construct the CIs by (a) computing the SE as proposed in Efron and Tibshirani (1997), (b) employing a confidence level of 99%, and (c) assuming that the prediction error estimates are distributed approximately normal. The runtime complexity of both bootstrap variants is $O(B * fitCost)$, where B is the number of bootstrap samples and $fitCost$ is the time complexity of estimating model parameters.

CM

The CM was proposed by Wagenmakers, Ratcliff, Gomez, and Iverson (2004) as a way to assess to what extent two models are able to mimic each other's behavior. Since model complexity and the ability to mimic other models are often related, the obtained mimicry information potentially allows reducing the bias towards selecting more complex models.

The following steps are involved in the CM: First, for one of the models (say, model 1) a certain number, *NDS*, of sets of parameter values are randomly drawn from the feasible range of the model's free parameters. Second, model 1 is used to generate *NDS* data sets employing each of the *NDS* parameter value sets, respectively. Third, both models are fitted to each of the *NDS* data sets yielding *NDS* measures of goodness of fit (GOF, e.g., the mean squared error) for both models. Fourth, the pairwise GOF differences are computed for all datasets. By repeating these four steps for the second model (model 2), one obtains two distributions of GOF differences, one for data generated from model 1 and one for data generated from model 2.

Given a set of observations, these two distributions can be utilized to decide which of the two models provides the better account of the observations. Both models are fitted to the observations and the difference in the models' GOFs are computed. If the resulting difference is classified to more likely come from the distribution resulting from data generated from model 1, model 1 is assumed to be more appropriate; otherwise the model 2 is assumed to be more appropriate.

Based on the results reported in Schultheis and Singhaniya (accepted), we employed a variant of the k-Nearest Neighbor algorithm ($k = 10$) for classification. The runtime complexity of the CM is $O(NDS * fitCost)$.

Simple Hold-Out

This method gauges the to-be-compared models by repeatedly splitting the set of available n observations into a training and test set. For each of these splits, both models are fitted to the respective training set. The fitted models are then used to generate predictions for the data points in the test set and the corresponding prediction error is determined. Accordingly, using I different splits results in I prediction error values for each of the two models. The model that has the lower median prediction error is selected as the more appropriate model. The runtime complexity of the simple hold-out method is $O(I * fitCost)$.

PED

Similar to the simple hold-out the PED (van de Wiel, Berkhof, & van Wieringen, 2009) employs I splits of the original data set into training and test set to compare models. For both models the prediction error is computed for each point in the test set after fitting the models to the corresponding training set. Subsequently, pairwise differences between prediction errors for model 1 and model 2 are calculated. These signed error differences are subjected to signed rank tests to derive the probability of the observed distribution of signed ranks under the null hypothesis that the models do not differ in predictive accuracy.

Thus, the PED yields I probability values. If the median of these values is below or equal to a pre-specified significance level α, the models are assumed to be significantly different in their predictive accuracy and the model with the smaller prediction error is assumed to be the more appropriate model. In our simulations we used the Wilcoxon signed rank test with $\alpha = 0.05$. The runtime complexity of the PED is $O(I * fitCost)$.

Method Properties

The procedural details of the methods described above imply a number of (differences in) crucial properties of the methods regarding model comparison.

First, the methods apply different criteria for judging the suitability of the compared models for a given data set. Both bootstrap variants, the PED, and the simple hold-out judge the models based on their ability to generalize to new data points, that is, these methods attempt to optimize what has

been called the *generalization criterion* (Busemeyer & Wang, 2000). In contrast, the CM has been argued to be optimal "under the validation criterion of selecting the generating model" (Cohen et al., 2008, p. 698). Since our simulations check the methods ability to recover the generating model, they test the conjecture of (Cohen et al., 2008) or, more generally, examine to what extent methods employing different criteria perform (dis)similarly in model recovery.

Second, only the bootstrap without SE and the simple holdout method can straightforwardly be extended to the simultaneous comparison of more than two models. All other methods are (currently) restricted to comparing pairs of models.

Third, the bootstrap with SE and the PED are the only methods that explicitly take into account the statistical variability and reliability during comparison. This renders these methods potentially superior to the other methods, because statistically reliable decisions between models can be assumed to be more accurate. On the other hand this property comes with the potential disadvantage that no decision may be possible in certain situations[1]. Accordingly, the overall quality of the bootstrap with SE and the PED will depend on the precise tradeoff between how accurately a decision between models can be taken and the number of situations in which a decision is reached.

Approach

Three hypothetical models of memory decay, $M1, M2$, and $M3$, were used to assess the model comparison methods. Each of these models predicts the probability of recall in dependance on the time t that has passed since the to-be-remembered items have been learned. The models are defined by the following formulas (see Pitt & Myung, 2002):

$$M1 : (1+t)^{-a},\ a \in [0,2]$$
$$M2 : (b+t)^{-a},\ a \in [0,2],\ b \in [1,2]$$
$$M3 : (1+bt)^{-a},\ a \in [0,2],\ b \in [0,2]$$

Note that $M1$ is nested in both $M2$ and $M3$, but nesting is different in the two cases. Since, furthermore, $M2$ and $M3$ are not nested, the three models allowed to examine the comparison methods regarding their ability to cope with different types of nesting as well as non-nested models.

Each method was applied to all three possible pairs of models, $M1$ vs. $M2$, $M1$ vs. $M3$, and $M2$ vs. $M3$ using the following general procedure. Given one of the three models, first, a set of parameter values was randomly drawn according to a uniform distribution from the range of parameter values specified above. Second, probabilities for this set of parameter values were generated from the model. Third, these probabilities were used to randomly sample the number of successful recalls from a binomial distribution assuming a certain number learned items (*NL*). Fourth, this set of numbers of successful recalls was treated as if it was a set of empirical observations for which to identify the most appropriate model. Accordingly, the comparison method in question was applied as described above to the model pair and the set of observations. Fifth, which (if any) of the two compared models was found to be more appropriate was noted. This procedure was repeated $R = 100$ times for each model in each model pair. Across all model pairs and methods the measure to assess model fits and prediction error was always the mean squared error and the models were fit using a variant of the Metropolis algorithm (Madras, 2002).

Following this general procedure, our simulations varied 5 factors that potentially impact the performance of the comparison methods. Besides allowing to assess the importance of each of these factors for method performance, factor variation ensured a more general view on the methods accuracy in model recovery, that is, a view that is not specific to only one particular combination of factor levels. The considered factors are tightness of fit, strength of noise, number of data points, number of samples, and split ration and are described in the following.

Tightness of fit Fitting a model to a set of observations is a specific instance of a general type of optimization problems: Find the optimal set of parameter values for the given observations. It is well known that one is rarely guaranteed to find the optimum in such optimization problems. Thus, model fits may often be suboptimal to greater or lesser extent. This raises the question how susceptible the different comparison methods are to suboptimal model fits. To investigate this, we considered 3 levels of tightness of fits by varying how thoroughly the Metropolis algorithm searches the models' parameter space. More precisely, we varied the number of sets of parameters that were sampled (called *swaps*) for model fitting, using swaps = 100, 1000, and 10000. Simulations looking at the rates for recovering the generating model when fitting to the probabilities directly (i.e., looking at model behavior without adding sampling noise) corroborated that these numbers of swaps realized increasingly accurate model fits.

Strength of noise Since the only noise in the data is sampling noise, the amount of noise in the data is determined exclusively by the number of learned items: The higher *NL* is the lower is the influence of sampling noise. Accordingly, employing $NL = 5, 50$, and 1000 allowed to examine the methods' capability to cope with noisy data.

Number of data points The information about the process that has generated a set of data can be assumed to increase with the number of available observations in the data set. To what extent the different methods require few or many data points for performing well was explored by varying the number of data points (*NDP*). Levels of $NDP = 5, 20$, and 100

[1] Some may also consider this a strong point of the methods, since the methods make explicit if too little information is available for a reliable decision. Yet, assuming that modelers often need to take a decision based on a set of available data, an equivocal comparison outcome is disadvantageous.

were employed and the corresponding data points were generated for t distributed equidistantly in the range $[0.1, 8.1]$.

Number of samples All of the methods come with a parameter that controls the amount of resources that are invested for model comparison. For PED and simple hold-out this parameter is the number of splits that are considered (I), for both bootstrap variants this parameter is the number of bootstrap samples (B), and for the CM this parameter is the number of GOF difference samples (NDS) each GOF difference distribution consists of. By using $I = 10, 100, 1000$, $B = 100, 1000$, and $NDS = 100, 1000$ we gauged the models resource-performance trade-offs.

Split ratio Application of the PED and the simple hold-out requires splitting the set of observations into training and test sets and the relative sizes of the two sets is potentially crucial for comparison performance. If the training set is too small, insufficient information about the generating process may be available. If the training set is too large, the danger of over fitting may arise and the test set may become too small to obtain a reliable estimate of generalization performance. In our simulations we investigated splits with $Q = 0.2, 0.4$, and 0.6, where Q indicates the fraction of the original observations that are used for the training set.

To assess the methods' ability to outperform less elaborate approaches to model comparison, our simulations comprise the *Akaike Information Criterion* (AIC, Akaike, 1973) as the sixth method and a seventh method that we term *simple recovery*. Following the same general procedure as described above, simple recovery compares models by only considering the GOF of each model on the given data set: The model that provides the tighter fit is assumed to be the more appropriate model. Simple recovery and AIC simulations involve the same variations of the factors tightness of fit, strength of noise, and number of data points as employed for the 5 more sophisticated methods.

Results

To characterize the methods' performance we computed, for each method, model pair, and situation, the sum of the percentages of cases in which both (a) a clear decision between the two models of pair could be taken and (b) the actually generating model was correctly recovered. If, for example, for the model pair M1-M2, M1 was correctly recovered 90% of the time and M2 was correctly recovered 43% of the time, the performance measure was computed to be $90 + 43 = 133$. Similarly, for BSSE and PED the percentages of cases where no model could be recovered with certainty was computed as the sum of the percentages of such cases for each of the two compared models.[2]. From the such obtained values the first,

[2]Given this procedure, BSSE and PED sometimes show both high performance and high percentages of situations where no model was recovered. Such a pattern indicates that the method in question only rarely recovered any model, but if it did, it was accurate

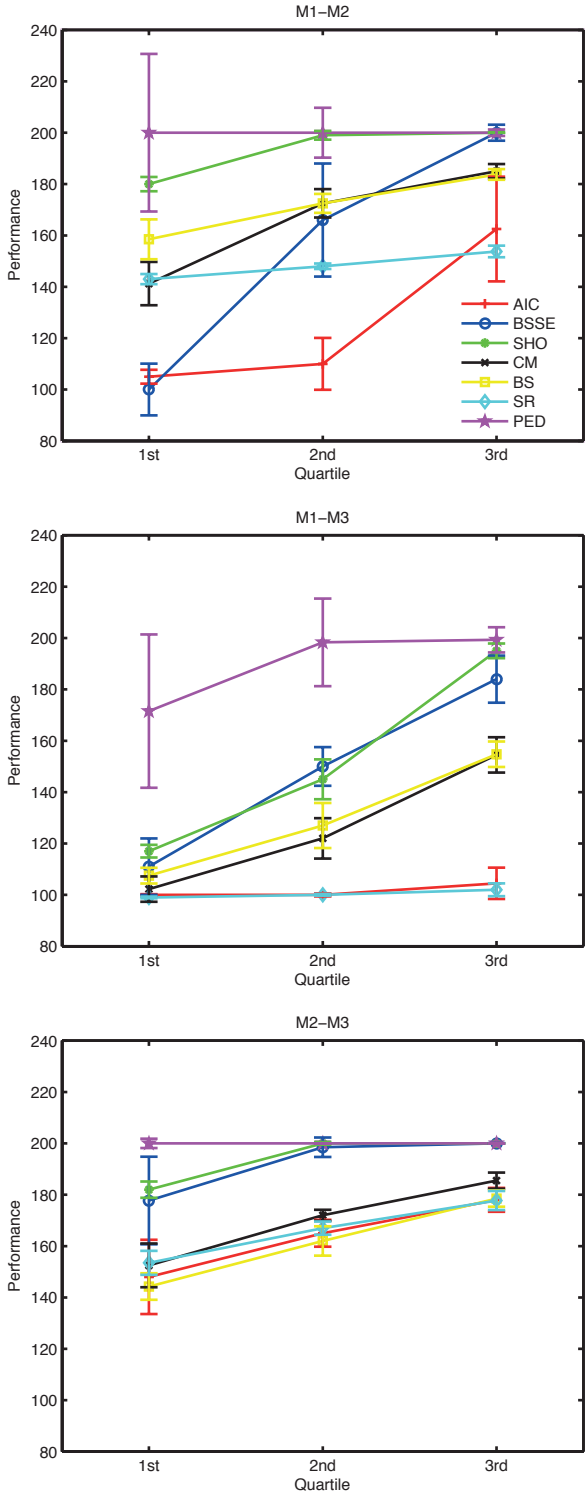

Figure 1: Quartiles of performance for the three considered model pairs and the seven considered methods. *AIC* = Akaike Information Criterion, *BSSE* = bootstrap with standard error, *SHO* = simple hold-out, *CM* = cross-fitting method, *BS* = bootstrap, *SR* = simple recovery, *PED* = PED method.

second (median), and third quartiles (and associated standard errors) were determined for each method and model pair across all situations. Figure 1 and Figure 2 display the quartiles for the different methods.

As is evident from Figure 1, there are marked performance differences between model pairs and comparison methods. As one may have expected, the nested model pairs generally prove more difficult than the non-nested model pair, with M1-M3 being even more difficult than M1-M2. It is mainly in the nested pairs that the less elaborate methods, AIC and simple recovery perform worse than all of the 5 more elaborate methods. Of the 5 more elaborate methods, PED, simple hold-out, and BSSE generally outperform BS and CM. In sum, PED, simple hold-out, and BSSE tend to perform best, AIC and simple recovery perform worst, and CM and BS show intermediate performance, but are only better than AIC and simple recovery for nested model pairs. As Figure 2 shows, the superior performance of PED and BSSE comes at the cost of a substantial number of cases in which the two methods do not allow to take a clear decision for one or the other model.

Several aspects of this pattern of results seem noteworthy. In contrast to the assumption that the CM is optimal for recovering the generating model (Cohen et al., 2008), the CM performs comparatively bad. On average, the CM is only better than SR for nested models, and generally worse in avoiding misclassifications than the PED, BSSE, and the simple hold-out. In fact, given its comparative simplicity, the simple hold-out performs remarkably well. While providing a decision for 100% of the cases, these decision are correct in more than 90% of the cases on average. This set of results also provides further evidence for dissimilarity in model recovery performance depending on whether a generalization criterion or a model recovery criterion is instantiated by the employed comparison method. Comparing the simple hold-out and CM indicates performance differences depending on which criterion is used and, more interestingly, that a method using the generalization criterion can outperform a method using the recovery criterion in model recovery.

In addition to the results across all factor combinations, considering the impact each factor has on method performance yields a number of interesting insights.

Tightness of fit Across all methods, the influence of the tightness of fit (if present at all) is only considerable between loose fits ($swaps = 100$) and moderate to tight fits ($swaps = 1000$ and 10000). In simple recovery, the tendency to select the more complex models increases with tightness of fits such that for moderate and tight fits the nesting model is selected more often even if the nested model generated the data. Except for pair M1-M3, performance of AIC increases considerably with tighter fits. In comparison, the bootstrap with SE and the CM, exhibit less (but still noticeable) susceptibility to tightness of fit in the sense that with tighter fits for nested model pairs the overall correct recovery rate increases by selectively increasing the correct recovery rate of

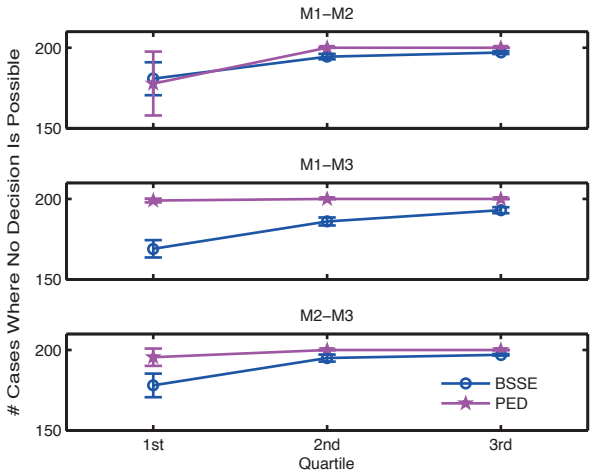

Figure 2: Quartiles of the number of cases for which *BSSE* and *PED* do not allow to take a decision.

the nesting model. Put differently, for loose fits, the CM and the bootstrap with SE tend to erroneously favor the simpler model; a problem that is mitigated when using tighter fits. The remaining three methods are largely insensitive to tightness of fits indicating that, for these methods, it may not be the absolute but the relative tightness of fit that matters.

Strength of noise Not surprisingly, all methods get consistently better with decreasing strength of noise. Furthermore, all methods encounter severe difficulties with the highest noise level ($NL = 5$) that leads to near chance performance for most model pairs and methods. The methods differ, however, regarding the level of noise from which they start to show good or very good performance. While the simple hold-out, the bootstrap with SE and the PED achieve high accuracy already for $NL = 50$, the CM and the bootstrap tend to do so only for $NL = 1000$.

Number of data points Although all methods but the AIC tend to improve with an increase in the number of data points, there are marked differences with respect to the strength of the influence of this factor. The PED and the bootstrap with SE are impacted severely by the number of data points improving considerably – especially for nested models – with an increase from $NDP = 5$ to $NDP = 20$ as well as from $NDP = 20$ to $NDP = 100$ both regarding accuracy and the percentage of decision that can be made. The other four methods are much less sensitive to *NDP* levels, but exhibit a tendency for a reduction in erroneously selecting a nested model when the data was generated from a nesting model. Interestingly, the performance of the AIC drops with increasing *NDP* due to an increased tendency to erroneously pick the nested model.

Number of samples Effects of increasing the number of samples are mixed across the methods. This factor has virtually no effect on the bootstrap. Yet, for the bootstrap with SE increasing the number of samples leads to a decrease in the percentage of cases in which a decision can be made and to a tendency to more often select the simpler of the two compared models. Both PED and simple hold-out perform better with increased I, but this trend is largely due to the difference between $I = 10$ and $I = 100$. Similar to the bootstrap with SE, the PED allows (slightly) fewer decision with increasing I. The CM exhibits a shift towards more often selecting the more complex model with increased numbers of samples.

Split ratio The split ratio has only little impact on the performance of the PED and the simple hold-out. While the number of cases that cannot be decided by the PED slightly increases with an increase in Q, the accuracy remains generally high. Only for comparing $M1$ and $M3$ do higher values of Q lead to pronounced performance decrements. Similarly, the simple hold-out becomes slightly but consistently worse in correctly recovering the nested model in the two nested model pairs with an increase in Q.

Conclusion

Our simulation studies revealed a number of interesting properties of the considered comparison methods. First, methods employing a generalization criterion for model comparison (e.g., simple hold-out) can outperform methods supposedly optimal for model recovery (the CM) in model recovery. Second, although all 5 considered methods can substantially improve on less elaborate approaches (as instantiated by the AIC and the simple recovery method), the less elaborate methods may perform better under certain conditions. Thus, whether the use of one of the examined methods is advantageous will depend on the precise nature of the model comparison situation at hand (e.g., how many data points are available and how noisy the data is). Third, the considered methods differ noticeably in the degree to which their performance depends on the characteristics of the comparison situation. The comparatively low quartiles of the bootstrap and the CM indicates that these methods outperform the less elaborate approaches only in comparatively few particular settings. Fourth, the highest accuracies were achieved by the PED, but this method allows decisions about which of the compared models is more appropriate only in very few cases. Furthermore, performance of the PED breaks down if only few data points are available. Fifth, despite its comparable simplicity, the simple hold-out method achieves high accuracies while allowing to select one of the models in 100% of all cases. In addition, the simple hold-out is the only method that can be easily extended to comparing more than two models.

Against this background our results suggest to employ the PED if only pairs of models have to be compared and if accuracy is more important than being able to reach a decision. The simple hold-out appears to be a good choice if more than two models need to be compared and / or if it is important to reach a decision on which of the compared models to select.

Although this initial assessment already highlights important properties of the comparison methods, it is best viewed as a first glimpse on the methods' characteristics. Further research considering a range of different (types of) models is required to provide a more comprehensive picture of the strengths and weaknesses of available comparison methods. Besides taking up this task we intend to explore modifications of the CM, PED, and bootstrap with SE that renders them applicable to comparing more than two models in our future work.

Acknowledgments

The authors gratefully acknowledge support by the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG) through the project R1-[ImageSpace], SFB/TR 8 Spatial Cognition

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov (Ed.), *Proceedings of the Second International Symposium on Information Theory* (p. 267 - 281).

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171-189.

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review*, *15*, 692-712.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Efron, B., & Tibshirani, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*, *92*, 548-560.

Madras, N. N. (2002). *Lectures on monte carlo methods*. Providence, Rhode Island: American Mathematical Society.

Pitt, M. A., & Myung, J. (2002). When a good fit can be bad. *TRENDS in Cognitive Sciences*, *6*, 421-425.

Schultheis, H., & Singhaniya, A. (accepted). Decision criteria for model comparison using cross-fitting. In *22nd Annual Conference on Behavior Representation in Modeling & Simulation (BRiMS 2013).*

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248-1284.

van de Wiel, M. A., Berkhof, J., & van Wieringen, W. N. (2009). Testing the prediction error difference between two predictors. *Biostatistics*, *10*, 550 - 560.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28-50.