Percentile analysis for goodness-of-fit comparisons of models to data

Sangeet Khemlani and J. Gregory Trafton
skhemlani@gmail.com, trafton@nrl.navy.mil

Navy Center for Applied Research in Artificial Intelligence
US Naval Research Laboratory, Washington, DC 20375 USA

Abstract

In cognitive modeling, it is routine to report a goodness-of-fit index (e.g., R^2 or RMSE) between a putative model's predictions and an observed dataset. However, there exist no standard index values for what counts as "good" or "bad", and most indices do not take into account the number of data points in an observed dataset. These limitations impair the interpretability of goodness-of-fit indices. We propose a generalized methodology, *percentile analysis*, which contextualizes goodness-of-fit measures in terms of performance that can be achieved by chance alone. A series of Monte Carlo simulations showed that the indices of randomized models systematically decrease as the number of data points to be fit increases, and that the relationship is nonlinear. We discuss the results of the simulation and how computational cognitive modelers can use them to place commonly used fit indices in context.

Keywords: goodness-of-fit, computational cognitive modeling, percentile analysis

Introduction

A common methodological practice for cognitive science researchers is to assess the merits of a cognitive model by evaluating its ability to capture the dynamics of a relevant dataset. For example, an adequate model of list memory might be one that captures appropriate serial position effects as well as other related psychological phenomena (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998). The predictions derived from such a model can themselves be tested experimentally. Thus, comparisons of theoretical predictions to empirical data reflect an alternating dialectic between theory building and experimentation (cf. McClelland, 2009).

A common way of assessing the fit of a model to data is to employ statistical goodness-of-fit measures. One such measure is the coefficient of determination (R^2), which is often interpreted as the proportion of variance explained by the model. Another is the root mean squared error (RMSE), a measure of the residuals between expected and observed values. Generally, R^2 is used to characterize the *precision* of a model, and RMSE is used to characterize a model's *accuracy* at accounting for a given dataset. They are often reported in concert with one another (Schunn & Wallach, 2005) under the assumption that a good model must be both precise and accurate. The indices are used ubiquitously as a method of model evaluation in computational cognitive science research (Busemeyer & Diederich, 2010). Indeed, in the *Proceedings the 35th Annual Conference of the Cognitive Science Society* (CogSci 2013) alone, 51 out of 171 papers (30%) self-identified as pertaining to computational modeling made use of at least one of the two indices to assess the fit of a cognitive model to empirical data. For instance, when Kachergis and Yu (2013) applied a computational model of cross-situational word learning (Kachergis, Yu, & Shiffrin, 2012) to an experiment they conducted, it revealed a high R^2 value (= .98), and the authors noted that their model "achieved quite a good fit to the data" (p. 713) and that it "could account for individuals' behavior in each of the conditions [of their experiment]" (p. 715). On the assumption that a good model must account for a generous proportion of the variance in a given dataset, R^2 provides a readily interpretable metric with which to evaluate and optimize computational models.

Our present analysis focuses on the limitations of using R^2 and RMSE as model evaluation metrics. Consider the fit of two hypothetical cognitive models, Theory A and Theory

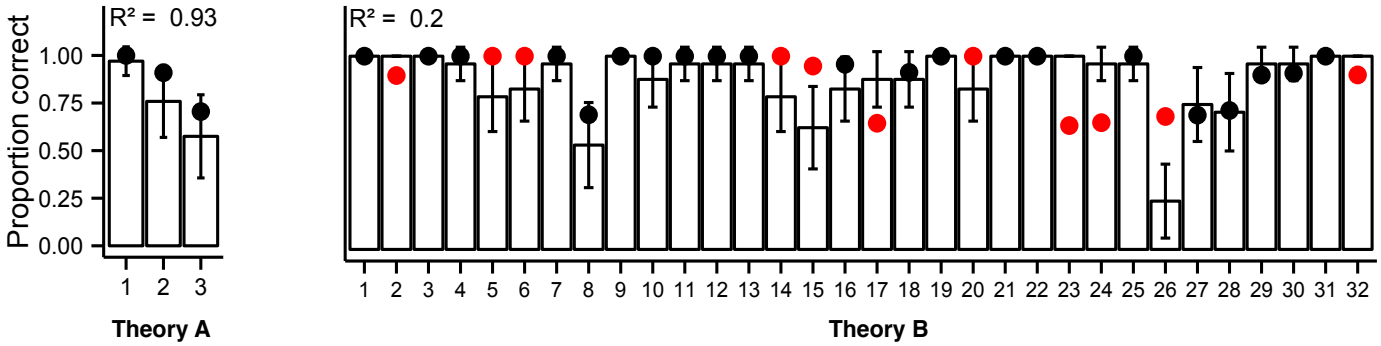

Figure 1. Hypothetical observed data (histograms and error bars) and hypothetical model predictions (circles) for two separate models, Theory A and Theory B. Errors bars reflect 95% confidence intervals. Black circles indicate when the predictions fell within the confidence interval of the observed proportion of correct responses, while red circles indicate deviations from the predictions and the observation. Theory A makes predictions of three separate problems, and Theory B makes predictions of thirty-two separate problems.

B, to hypothetical datasets as depicted in Figure 1. Theory A is conservative: it attempts to model the proportion of correct responses to three separate problems, and it accounts for 93% of the empirical variance. Theory B is more ambitious: it attempts to model the responses of thirty-two problems, but accounts for only 20% of the variance. Our intuitions might therefore suggest that Theory A provides a close match to the data, and Theory B, for all its ambition, provides a poor match to the data. The goal of the present article is to show how both intuitions can be incorrect: Theory B, we argue, reflects an excellent account of the data, and Theory A's fit is not particularly impressive. To show how this is the case, we turn to address several limitations of common model fitting metrics and discuss those limitations in the context of developing computational cognitive models. We introduce a novel, theoretically motivated metric, *percentile* R^2, which overcomes limitations of orthodox goodness-of-fit measures. We then describe a simulation study that reveals the dynamics of the percentile R^2 metric relative to different sorts of hypothetical datasets. Finally, we conclude by addressing constraints of the percentile R^2 metric and prescribe its use in computational cognitive modeling.

The limitations of goodness-of-fit

The use of goodness-of-fit measures like R^2 and RMSE is controversial: some researchers argue that they are uninformative and should not be used (Roberts & Pashler, 2000) while others suggest that the metrics themselves are informative, but they are often misused (Schunn & Wallach, 2005; Stewart, 2006). Both sides agree that a major problem with goodness-of-fit metrics is that there exist no established measures for how to interpret them (Estes, 2002): researchers often rely on conventions, such as that an $R^2 > .90$ reflects a "good" fit, and these conventions can be misleading. Furthermore, Schunn and Wallach acknowledge several other factors that affect the quality of a fit, including the noise in the data and its information-theoretic complexity.

One of the primary reasons that the conventions for good and bad fits are misleading is because R^2 does not reflect the number of data points (N) that a model attempts to fit. In other words, we might intuitively believe that a model that accounts for 95% of the variance amongst twenty data points is stronger than an alternative model that accounts for the same mount of variance amongst only five data points, provided that the data points are distributed in a non-trivial way, e.g., in a non-linear fashion. However, there is nothing inherent about the goodness-of-fit metrics themselves that rewards the fitting of more points (or penalizes the fitting of fewer points). There exist metrics for model selection that combine goodness-of-fit and model complexity measures, such as minimum description length (Grünwald, 2001), the Akaike information criterion (AIC; Akaike, 1973; Bozdogan, 2000), and the Bayesian information criterion (BIC; Schwarz, 1978; Wasserman, 2000), but these metrics generally take into account a model's free parameters in assessing its complexity, and not the number of data points it attempts to fit.

The insensitivity of goodness-of-fit metrics to the number of data points being fit may appear innocuous at first blush. After all, a model that accounts for 90% of the variance intuitively strikes us as an adequate account of the data regardless of how many points it fits. However, the established conventions for what is considered a "good" model, combined with the insensitivity to the number of data points under consideration, may discourage cognitive modelers to fit their model at the level of individual items and problems. If a computational model makes quantitative predictions for twenty separate items, but those individual items can be collapsed into four-item sets, the modeler may be tempted to optimize the model's fit to the set-wise analysis (five separate data points) and to ignore or else not report the item-wise analysis (twenty separate data points). Indeed, we would venture that many cognitive modelers would reject our hypothetical Theory B (see Figure 1) on the basis of how low the R^2 is, disregarding how many points the theory attempts to fit. To do so would be a failure to recognize that the number of data points is negatively correlated with the chance probability of obtaining a high R^2.

One solution to the problem is to consider a metric that is both sensitive to the number of data points under investigation as well as uniformly interpretable and meaningful. The next section introduces such a metric.

Percentile analysis

We posit a general methodology of model evaluation, *percentile analysis*, which contextualizes goodness-of-fit indices in terms of performance that can be achieved by chance alone. It is sensitive to the number of data points under investigation, as well as interpretable. The methodology is designed to take into account the strengths of orthodox goodness-of-fit metrics like R^2, which are often used under the assumption that a model is "good" whenever its predictions explain a large proportion of the data. The term "large" is usually taken to mean upwards of 90%, but it is merely a convention, and many studies attribute good fits to models whose R^2 values are lower (e.g., Lassiter & Goodman, 2012; Salvucci, 2005) depending on the particular dataset. The threshold for evaluating R^2 is therefore ambiguous. A less ambiguous assumption of model fitness can be achieved by comparing the model to hypothetical alternatives. That is, the metric we propose, *percentile* R^2, assumes that an acceptable model is one that accounts for more variance than a set of predictions produced by chance alone, and it can be scaled and interpreted based on the probability that the results could occur by chance.

As an illustration, consider an observed dataset, $D_{OBSERVED}$, that consists of N data points that one wishes to model, and a set of predictions as derived from a putative model, $D_{PREDICTED}$. Our goal is to describe how well the predicted data fits the observed data. We do this by

exploring the space of possible models as produced by chance by generating randomized models, i.e., sets of *N* data points at random, D_1 to D_{100}. We then calculate the proportion of explained variance, R^2, for $D_1 \ldots D_{100}$ and for $D_{PREDICTED}$ relative to $D_{OBSERVED}$. Finally, we examine the percentile rank of our putative dataset, $D_{PREDICTED}$, relative to the 100 randomized models. The percentile rank is what is reported as the percentile R^2, and we accordingly define it as follows:

$$Percentile\ R^2 = \left(\frac{f_{below} + \frac{1}{2} f_{within}}{N} \right) * 100$$

where:

- f_{below} is the frequency of random models whose R^2 values are less than the R^2 value of the putative model
- f_{within} is the frequency of random models whose R^2 values are the same as the R^2 value of the putative model
- N is the number of random models.

Thus, a percentile R^2 of .93 means that the putative model's R^2 value is higher than 93% of the R^2 values of the random models. Clearly, the fidelity of the percentile R^2 measure depends on the number and quality of the datasets generated by the random exploration of the data space. Nevertheless, the law of large numbers guarantees convergence on a stable percentile R^2 as the sample size of random models increases.

The percentile R^2 is advantageous because it allows researchers to uniformly evaluate theories that differ in the number of predictions they make on their ability to account for a given data set. Suppose one model (Model A) makes fewer predictions than another model (Model B). Employing percentile R^2 as a metric in concert with R^2 to evaluate Model A and Model B has four potential outcomes:

1. *Model A has a higher R^2 value than Model B, and also has a higher percentile R^2 than Model B.* In this uncontroversial case, Model A is universally preferred over Model B. Assuming that the model's R^2 is high enough to account for an large proportion of the data relative to the standards set by other researchers, it is declared to have a fit that is both sufficiently high and sufficiently ambitious, i.e., it fits that data better than its competitor.

2. *Model A has a higher R^2 value than Model B, but the two models have identical percentile R^2 values.* In many cases, this scenario is uncontroversial: in the event that percentile R^2 values between two theories match, the model with the higher R^2 is interpreted as fitting the most data. However, it is possible that Model A has only a marginally higher R^2 value relative to Model B (e.g., .95 vs. .92). In this event, the two models have comparable accounts of the data, i.e., they cannot be distinguished on the virtue of model fits alone.

3. *Model A has a higher R^2 value than Model B, but a lower percentile R^2 value than Model B.* In this controversial case, if R^2 is used as the only goodness-of-fit metric, then Model A will be deemed to have a closer fit to the data than Model B, disregarding the fact that Model B makes more predictions and is susceptible to a lower fit by virtue of chance alone. In contrast, taking percentile R^2s into account yields one of two separate conclusions: either Model A is deemed to be not sufficiently ambitious, i.e., not able to account for as many data points as Model B, or else Model B is deemed to have poor descriptive power (although it may be more generalizable than Model A; see Cavagnaro, Myung & Pitt, 2013). The key insight of this controversial scenario is that a cognitive modeler is not justified in dismissing Model B on account of its inability to account for enough variance, and must direct criticisms to other facets of the model (e.g., its parsimony, breath, and ability; see Cassimatis, Bello, & Langley, 2008).

4. *Model A and Model B have identical R^2 values relative to the datasets they attempt to fit.* In this case, Model B is guaranteed to have a higher percentile R^2 value than Model A, and is therefore deemed to have a closer fit to the data than its competitor.

In the first case, the use of percentile R^2 reinforces the results of employing the orthodox goodness-of-fit metric, R^2, alone. The latter cases, however, are those that pose a challenge for metrics that disregard the number of data points being fit. In those cases, percentile R^2 provides meaningful, contextually relevant interpretations of model fits, and allows a modeler to adjudicate between putative models.

To explore these latter two cases, we conducted a Monte Carlo simulation of observed datasets ($D_{OBSERVED}$) of varying numbers of items in the dataset (*N*), and calculated the R^2 and RMSE values at informative percentiles. We expected that R^2 would drop and RMSE would rise relative to *N*. Our goal, however, was to examine these patterns as well as the raw goodness-of-fit values that achieve relatively low and relatively high percentile R^2s.

Monte Carlo simulation study

We conducted a series of Monte Carlo simulations to explore the distribution of percentile R^2 values as a function of the number of items in the dataset. Random samples were drawn to numerically investigate the properties of the unknown probability distribution defining percentile R^2 values.

Method and procedure. 127 separate Monte Carlo simulations were conducted for values of N ranging from 2 to 128. Each of the simulations was run 10,000 times. Every run comprised three operations:

1. First, an *observation* sample, $D_{OBSERVED}$, is defined by drawing N samples from a standard uniform probability distribution, $U(0, 1)$.

2. A *model* sample, D_i, is then constructed by again drawing N samples from $U(0, 1)$.

3. Goodness-of-fit statistics, i.e., R^2 and RMSE, are calculated between D_i and $D_{OBSERVED}$.

The simulations drew samples from the uniform probability distribution, which ranges from 0 to 1. Such a distribution can be used to model proportions, e.g., accuracy data. We performed an analogous simulation study by drawing samples from a unit normal distribution, but as its results were largely similar to those of the analysis performed over the uniform distribution, we omit them for brevity. However, the second simulation made evident that the distribution from which samples are drawn makes a substantive difference in the analysis of RMSE and other metrics of deviation from exact location.

After the simulation was carried out for a given value of N, the system calculated the values of R^2 and RMSE at four separate percentiles of salience. Three of the percentiles, the 90th, 95th, and 99th, represent those that correspond to orthodox alpha values in inferential statistics, i.e., they represent percentiles that could potentially be deemed a "good" fit. Values of R^2 and RMSE at the 99th percentile, for example, indicate that of 100 random guesses, on average only one will match or exceed it. A fourth percentile value, the 70th percentile, provided a value of what would unequivocally be considered a "poor" fit. In other words, it described the values of R^2 and RMSE that could be achieved (or surpassed) 30% of the time if a theory structured its predictions at random.

Results of the simulation

The Monte Carlo simulation provided a numerical exploration of the multivariate probability distribution that defines percentile R^2 values. Figure 2 plots average R^2 and RMSE values as a function of the four percentiles analysed for values of N that range from 2 to 100. Table 1 provides mean R^2 and RMSE numerical values at the four percentiles of interest for values of N at powers of 2. The results of the simulations reveal the predicted monotonic trends: R^2 values drop as the number of data points in a model increases while RMSE values increase proportionally to the number of data points in a model. These results serve as manipulation checks and validate the fidelity of the simulation.

As Figure 2 shows, the simulations revealed that when fitting low numbers of data points, a set of random predictions could achieve high R^2 values. For instance,

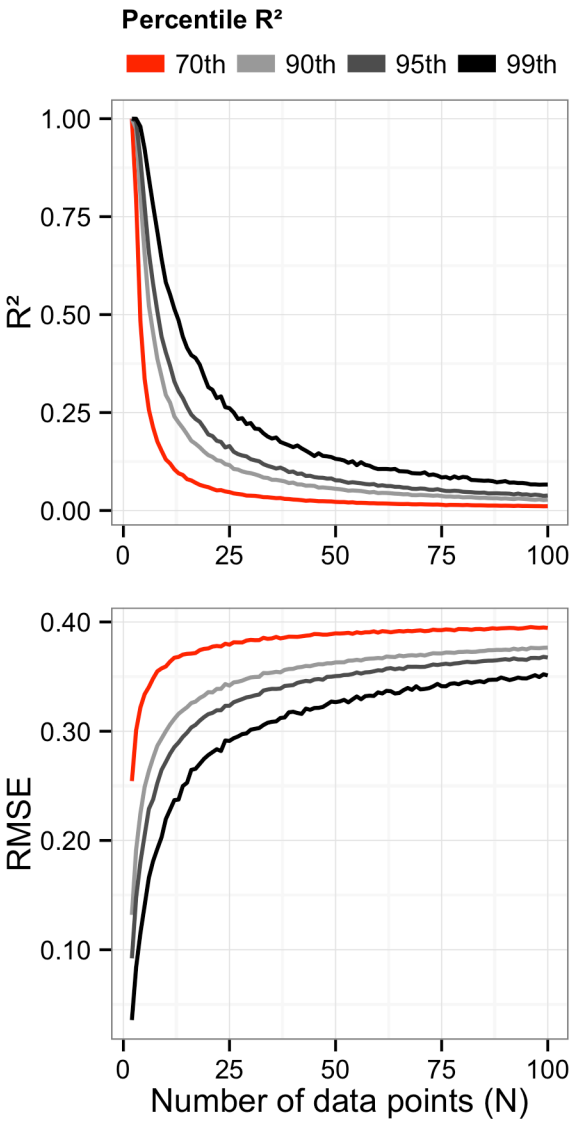

Figure 2. Mean R^2 (top panel) and RMSE (bottom panel) values at the 70th, 90th, 95th, and 99th percentiles of R^2 values where the number of data points in the model (N) ranges from 2 to 100 (truncated from 128).

when fitting 4 data points, around 10% of random predictions (i.e., those at the 90th percentile or higher) yield R^2s $\geq .97$. That is, one in ten random guesses can capture nearly all of the variance among 4 data points. Increasing the number of data points to 8 makes it progressively more difficult to achieve a high R^2 value by chance alone: the model at the 99th percentile yielded an $R^2 = .71$. By the time a modeler attempts to fit 100 data points, the probability of achieving a high R^2 by guessing randomly is astronomically low, and the model at the 99th percentile yielded a paltry R^2 of .06.

The results of the simulation study provide the basis of a new measure of goodness-of-fit that is imminently interpretable. They yield a systemized account of what can be considered an acceptable fit of the data. For example, a

	R^2				RMSE			
	Poor	*Good*			*Poor*	*Good*		
N	70th	90th	95th	99th	70th	90th	95th	99th
2	1.00	1.00	1.00	1.00	0.22	0.12	0.09	0.04
4	0.49	0.85	0.93	0.99	0.38	0.26	0.21	0.13
8	0.17	0.39	0.51	0.72	0.34	0.27	0.24	0.18
16	0.08	0.18	0.25	0.39	0.38	0.33	0.31	0.27
32	0.04	0.09	0.12	0.20	0.37	0.34	0.33	0.30
64	0.02	0.04	0.06	0.10	0.41	0.39	0.38	0.36
128	0.01	0.02	0.03	0.05	0.40	0.38	0.38	0.36

Table 1. Mean R^2 and RMSE numerical values at the four percentiles of interest for values of *N* at powers of 2.

model's alleged "good fit" (R^2 = .95, RMSE = .30) can be corroborated or contravened depending on whether its percentile R^2 is .80 or .99. We conclude by discussing the broader use of percentile analysis in computational cognitive modeling.

General Discussion

A good fit of a model to data can promote refinement; a bad fit can encourage correction or dismissal. The analytical technique we propose and implement, *percentile analysis*, augments traditional goodness-of-fit measures with information regarding how well a model performs relative to what can be achieved by chance alone. A series of Monte Carlo simulations provide the numerical basis for novel metrics that can be used to place model fits in context. For example, consider the fits of the two theories we introduced at the beginning of the paper, Theory A and Theory B (see Figure 1). Theory A appears to have a closer fit to the data (R^2 = .93), and despite the fact Theory B makes more predictions, it appears to achieve a poor fit to the data (R^2 = .20). Percentile analysis provides the full context: the percentile R^2s of Theory A and Theory B are .81 and .99 respectively. That is, on average, roughly one in five random predictions can achieve a fit that matches or exceeds Theory A's performance, whereas less than one in a hundred random predictions can exceed Theory B's performance.

The preceding example demonstrates how percentile analysis can promote better computational modeling practices. Because fit indices are guaranteed to drop as the number of data points increases, we suspect that many computational modelers are susceptible to Type II error: they disregard models of potential value because they understand that an R^2 of .20 is unlikely to impress reviewers. Percentile analysis provides a more thorough perspective for evaluating models: a model whose R^2 is .20 may be worth considering on the basis of how difficult it would be to account for 20% of the variance by chance alone.

Perhaps a more pressing concern is that modelers are susceptible to Type I error as well: a purported "good" fit (e.g., R^2 = .93) might be a result of a random allocation of predictions. In this case too, percentile analysis allows for a motivated rejection of the model: if the percentile R^2 value is low, then its predictions may be dubious.

In either event, the use of percentile R^2 in computational modeling promotes the analysis of detailed model fits. Cognitive modelers often fit their models at the level of sets of data points by aggregating individual problems and items in some theoretically meaningful way. They rarely fit their models at the level of individual items, and it is no surprise: modelers who do so are often guaranteed to yield what orthodox goodness-of-fit metrics would consider a "poor" fit. Nevertheless, a good model should be able to fit the data at *both* the set-wise level and the level of individual items (provided that the data collection methodology is robust). Percentile analysis is a tool that allows for the development of such models.

Percentile analysis can also aid in more common computational modeling practices, such as parameter optimization. R^2s and percentile R^2s are separate measures, i.e., the latter is a function of the number of items in a dataset, whereas the former is not. As a result, percentile R^2s may prove to be a better index of fitness with which to tune parameter settings, and optimizing for percentile R^2s can potentially find parameters that avoid overfitting. That is because, for sufficiently large values of *N*, the parameter space that yields percentile R^2s > .90 is larger than the space that yields R^2s > .90, and the additional parameter settings may allow for more generalizability. The end result could be a parameterization of a model that yields a relatively "low" R^2 value (< .90), a high percentile R^2 value, and better cross-validation potential.

Are there disadvantages to employing percentile analysis in computational modeling? Critics may wonder if the introduction of yet another index of model fit is worthwhile: they might hold that present methodologies suffice to quantify how much variability a models accounts for, and

that metric alone serves as a sufficient metric for evaluating models. However, metrics such as R^2 and RMSE suffer from many limitations as reviewed above, not the least of which is that they are difficult to interpret. As a poor model fit is regarded in such disdain as to present an impediment to publication, we argue that percentile analysis is an indispensible tool for interpreting goodness-of-fit indices and placing them in an appropriate context. Percentile analysis is not a methodology meant to supplant orthodox goodness-of-fit measures, but rather one that should be used to make them more comprehensible. The same points hold mutatis mutandis for model selection metrics like minimum description length, AIC, and Bayesian non-parametric approaches (Karabatsos, 2006): the advantage of these metrics is that they take into account model complexity, but their disadvantage is that they are hard to interpret and to compare across datasets.

Dissenters may also hold that percentile analysis is a vindication for poor modeling: it allows for the publication of models that account for relatively little variance in the data (e.g., R^2 = .20). Far from the dissenting position, however, we believe percentile analysis promotes better computational modeling practices, because it contextualizes previous methodologies and allows modelers to reasonably examine detailed model fits across individual items. Therefore, a modeler's focus needn't rest on maximizing R^2 values alone; they can also try to build models that are flexible enough to make detailed process predictions.

In summary, we developed percentile analysis as a methodology for evaluating and interpreting a model's fit to observed data. It is sensitive to the number of data points in the data, and it operates by comparing a putative model against hypothetical ones generated by pure noise. We argue that percentile analysis should be an essential component of a cognitive modeler's toolkit.

Acknowledgements

This research was funded by a National Research Council Research Associateship awarded to SK and a grant from ONR awarded to GT. We are also grateful to Bill Adams, Magda Bugajska, Dan Gartenberg, Tony Harrison, Laura Hiatt, Ed Lawson, Frank Tamborello, and Alan Schultz for their helpful comments.

References

Akaike, H., & (Eds.). (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Kiado, Budapest: Akad.

Anderson, J. R., Bothell, D., Lebiere, C. & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, 38, 341-380.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in statistics theory and methods,* 19, 221-278.

Busemeyer, J. R. & Diederich, A. (2010). *Cognitive Modeling*. Sage.

Cassimatis, N., Bello, P. & Langley, P. (2008). Ability, breadth and parsimony in computational models of higher-order cognition. *Cognitive Science,* 32, 1304-1322.

Cavagnaro, D. R., Myung, J. I., & Pitt, M. A. (2013). Mathematical modeling. In Todd D. Little (ed.), *The Oxford Handbook of Quantitative Methods*, Vol. 1 (pp, 438-453), Oxford University Press, New York, NY.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin &Review,* 9, 3-25.

Grünwald, P. (2001). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44, 133-152.

Kachergis, G. & Yu, C. (2013). More naturalistic cross-situational word learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). Cross-situational word learning is better modeled by associations than hypotheses. *IEEE Conference on Development and Learning / EpiRob 2012.*

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology,* 50*,* 123-148.

Lassiter, D. & Goodman, N. D. (2012). How many kinds of reasoning? Inference, probability, and natural language semantics In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*.

McClelland, J. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science,* 1*,* 11-38.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358-367.

Schunn, C., & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In W. Tack (Ed.), *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115-154). University of Saarland Press, Saarbrücken, Germany.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* 6, 461-464.

Stewart, T. C. (2006). Tools and techniques for quantitaive and predictive cognitive science. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 816-821). Vancouver, British Columbia, Canada.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology,* 44, 92-107.