An Efficient Method for the Minimum Description Length Evaluation of Deterministic Cognitive Models

Michael D. Lee (michael.lee@adelaide.edu.au)
Department of Psychology, University of Adelaide
South Australia, 5005, AUSTRALIA

Abstract

The ability to evaluate competing models against noisy data is central to progress in cognitive science. In general, this requires advanced model selection criteria, such as the Minimum Description Length (MDL) criterion, that balance goodness-of-fit with model complexity. One limiting property of many of these criteria, however, is that they cannot readily be applied to deterministic models. A solution to this problem, developed by Grünwald (1999), involves a process called 'entropification' that associates deterministic models with probability distributions, and allows MDL criteria to be calculated. However, a potential practical difficulty with this approach is that it requires a multidimensional summation over the data space that can be prohibitively computationally expensive in realistic situations. This paper derives a simpler version of the MDL criterion for deterministic models in the important special case of 0-1 loss functions that is computationally feasible. Two concrete applications of the simpler MDL criterion are presented, demonstrating its ability to consider model fit and complexity in selecting between competing models of cognitive processes. The first application involves three different heuristics for a problem solving task, while the second involves three different models of forced-choice decision making.

Introduction

To a large extent, progress in cognitive science relies on the development of better models of cognitive phenomena. Models provide a formalized representation of theoretical explanations, and make predictions that can be tested empirically. For this reason, the ability to evaluate competing cognitive models against noisy data in a complete and meaningful way has been a central concern recently in mathematical psychology (e.g., Myung & Pitt 1997; Myung, Balasubramanian & Pitt 2000; Myung, Forster, & Browne 2000; Pitt, Myung, & Zhang 2002).

In particular, there has been a strong (and overdue) focus on balancing the goodness-of-fit of models with their complexity. These ideas have been applied to core topics in cognitive science such as models of psychophysical discrimination (e.g., Myung *et al.* 2000), stimulus representation (e.g., Lee 2001; Navarro & Lee 2003; in press), inference and generalization (e.g., Tenenbaum & Griffiths 2001), and decision-making (e.g., Myung & Pitt 1997).

Probabilistic Models

For the most part, however, these recent development have been restricted to considering *probabilistic* cognitive models. This class of models has the property that any parameterization (or, more generally, any probability distribution over the parameter space) corresponds to a probability distribution over the data. That is, the model corresponds to a parametric family of probability distributions over the data. This means that considering a probabilistic model at a particular set of parameter values makes some data quantifiably more likely than others. In turn, for probabilistic models the likelihood of any observed data having arisen under the model at any parameterization of interest can be evaluated.

Many cognitive models are probabilistic in this way. For example, models of memory retention (e.g., Rubin & Wenzel 1996) usually consist of parameterized functions that specify the probability an item will be recalled correctly after a period of time. As another example, the ALCOVE model of category learning (Kruschke 1992) also produces a probability, that depends upon the values of a number of parameters, for each possible category response on any trial. For these models, their probabilistic nature allows likelihood to be measured against any pattern of observed data.

Many advanced model selection criteria, such as Bayes Factors (e.g., Kass & Raftery 1995), Minimum Description Length (MDL: e.g., Grünwald 2000), Stochastic or Geometric Complexity (Myung, Balasubramanian & Pitt 2000; Rissanen 1996), and Normalized Maximum Likelihood (Rissanen 2001), rely on this property. This is because they integrate the probabilities of the data across the parameter space of the models, or the maximum likelihoods across all possible data sets, and so require non-zero probabilities over a subset of the parameter space that has measure greater than zero to be meaningful.

Deterministic Models

As Myung, Pitt and Kim (in press) note, however, there are many important cognitive models that belong to the alternative class of *deterministic* models. These models specify differently how to assess the relationship between data on the one hand, and model predictions at different parameterizations on the other. For example, a sum-squared loss or error function might be proposed, so that increasingly large differences between model predictions and observed data are penalized more heavily in evaluating the model. Alternatively, a 0-1 loss function might be proposed, so that models are evaluated as being correct only if they predict data exactly, and are wrong otherwise. What deterministic models do not specify, however, is an error theory that describes the likelihood of data that differ from model predictions. This means that, when a deterministic model makes incorrect predictions, it is not possible to assign the probabilities needed by many modern model selection criteria.

A good example of a deterministic cognitive model is the 'Take the Best' model of decision making (Gigerenzer & Goldstein 1996). This model takes the form of a simple algorithm, searching a fixed stimulus environment in a deterministic way, so that it will always make the same decisions. One way of interpreting the model in relation to empirical data is that it has probability one when it makes the same decision as that observed, but probability zero when it makes a different decision. Adopting this approach, however, any evaluation of the model against human data involving multiple decisions is very likely to find an overall probability of zero, because at least one of the model's decisions will disagree with the data.

Other deterministic models that face similar problems include the memory models surveyed by Pietsch and Vickers (1997), axiomatic theories of judgment and choice (e.g., Luce 2000), and various lexicographic decision models (e.g., Payne, Bettman & Johnson 1990). For these sorts of models, the natural assessment is in terms of the proportion of correct decisions it makes, or some such error function, but this measure is not the same as the probabilities from likelihood functions used in probabilistic model selection. In particular, it is not clear how the error function measuring goodness-of-fit should be combined with measures of model complexity to undertake model selection.

Recently, however, Grünwald (1999; see also Myung, Pitt, & Kim in press), has developed a model selection methodology that overcomes these difficulties. He provides a principled technique for associating deterministic models with probability distributions, through a process called 'entropification', that allows MDL criteria for competing models to be calculated. There is a potential practical difficulty, however, in using this approach to evaluate cognitive models. The MDL criterion involves multidimensional summations over the data space that could be prohibitively computationally expensive in some realistic situations. This paper derives and demonstrates a reformulation of the MDL criterion for deterministic models in the important special case of 0-1 loss functions that is much less computationally expensive.

The MDL Criterion

In this section, Grünwald's (1999) formulation of the MDL criterion based on entropification is described, and a computationally simpler form is then presented. In one sense, the reformulation is just a straightforward algebraic manipulation, and has probably been noted (but not published, as far as we are aware) by others. In another sense, making the reformulation explicit, and demonstrating its advantages, is a useful contribution. There are many cognitive models that are deterministic and naturally assessed under 0-1 loss[1], for which the MDL method described here ought to find wide application.

Original Formulation

Suppose a deterministic model M is being evaluated using a dataset D that has n observations, $D = [d_1, \ldots, d_n]$. Each of the observed data are discrete, and can assume only k different values. The model uses P parameters $\theta = (\theta_1, \ldots, \theta_P)$ to make predictions $Y = [y_1, \ldots, y_n]$. To evaluate any prediction made by the model, a 0-1 loss function is defined as $f(D, Y) = \sum_{i=1}^{n} \gamma_i$, where $\gamma_i = 0$ if $d_i = y_i$ and $\gamma_i = 1$ otherwise. By considering all possible parameterizations, the model makes a total of N different predictions. In other words, there are N different predictions, $Y_1, \ldots, Y_N$, the model is able to make about the data by choosing different parameter values. In general, the relationship between parameterizations and predictions will be many-to-one. This means that every unique model prediction is naturally associated with one or more parameterizations of the model.

Under these assumptions, Grünwald (1999) shows that using entropification the model making prediction Y can be associated with a probability distribution, parameterized by the scalar w, as follows:

$$p(D \mid M, Y, w) = \frac{e^{-wf(D,Y)}}{\sum_{x_1=1}^{k} \cdots \sum_{x_n=1}^{k} e^{-wf(D,[x_1,\ldots,x_n])}}.$$

Determining the MDL criterion for the model requires finding the model predictions Y^* and scalar w^* that *jointly* maximize $p(D \mid M, \theta, w)$ to give the value p^*.

[1] All of the deterministic decision making, memory and judgment models already mentioned effectively have 0-1 loss when they are restricted to two choices. There are other models, such as the optimal stopping models considered later, that are also naturally associated with 0-1 loss despite having a larger number of choices.

Once this is achieved the MDL criterion for the model is given simply by $\mathrm{MDL} = -\ln p^* + \ln N$.

Besides automatically balancing the competing demands of model fit and complexity, this MDL criterion has at least two attractive properties for model selection in cognitive science. First, differences in MDL values, through their natural probabilistic interpretation, can be assessed as odds, in much the same way as Bayes Factors. This allows the assessment the 'significance' of different MDL values for different models to be done meaningfully as a question of the standards of scientific evidence required for the problem at hand, using a scale that is calibrated by betting. Secondly, as Grünwald (1999, pp. 24-28) discusses, the information theoretic or coding approach used by MDL means that results are available for cases where the data generating process that is being modeled has statistical properties that are not perfectly represented by the models being considered. We would argue this is inevitably the case for cognitive models, and so the ability of the MDL approach to address this problem is an important one.

Despite these attractions, however, there is an obvious difficulty in maximizing $p(D \mid M, \theta, w)$. The problem is that the denominator given by $Z = \sum_{x_1=0}^{k} \cdots \sum_{x_n=0}^{k} e^{-wf(D,[x_1,\ldots,x_n])}$ involves considering every possible data set that could be observed, which involves a total of k^n terms. In cognitive science, where it is possible for a deterministic model to be evaluated using many data points, each of which can assume many values, the repeated calculation of Z may be too computationally demanding to be practical.

A Simpler MDL Computation

A simpler form for Z can be derived by noting that $f(D,Y)$ can only take the values $0,\ldots,n$, in accordance with how many of the model predictions agree with the data. Since Z considers all possible data sets, the number of times $n-x$ matches (i.e., x mismatches) will occur is $\binom{n}{x}(k-1)^x$. For a prediction Y that has $n-m$ matches with the data (i.e., there are m mismatches and $f(D,Y)=m$), this leads to the simplification

$$p(D \mid M, Y, w) = \frac{e^{-wm}}{\sum_{x=0}^{n} \binom{n}{x} (k-1)^x e^{-wx}},$$

which has a denominator that sums $n+1$ rather than k^n terms.

The computational efficiency offered by this reformulation means it will generally be possible to find the w_i^* that maximizes $p(D \mid M, Y_i, w_i)$, giving p_i^*, for all N model predictions. The p^* required for MDL calculation is then just the maximum of $p_1^*, \ldots, p_N^*$.

Finding each w_i^* can also be done efficiently by observing that

$$\partial p / \partial w = \frac{1}{Z^2} e^{-wm} \sum_{x=0}^{n} \binom{n}{x} (k-1)^x (x-m) e^{-wx}.$$

This derivative is clearly always positive if $m=0$ and always negative if $m=n$. This means, if a model predicts all of the data correctly, $w_i^* \to \infty$, and if a model fails to predict any of the data correctly $w_i^* \to -\infty$. Otherwise, if $0 < m < n$, the substitution $u = e^{-w}$ allows w_i^* to be found from the positive real roots of the degree n polynomial

$$\sum_{x=0}^{n} \binom{n}{x} (k-1)^x (x-m) u^x.$$

by standard numerical methods (e.g., Forsythe, Malcolm, & Moler 1976).

Grünwald (1999, pp. 98-99) notes, with particular reference to the 0-1 loss function, that the case $w < 0$ corresponds to 'inverting' models. For example, if a model only makes two choices, and so considers binary data (i.e., $k = 2$), the inverted model changes all of the model predictions to the alternative possibility. We would argue it will generally be the case in cognitive modeling that it is not appropriate to consider inversion, because this manipulation will require the model to be interpreted in a substantively different and unintended way. If this is the case, it is necessary to restrict consideration to $w \geq 0$ in finding the MDL value.

With this restriction in place, the Y^* and w^* learned from data for qualitative model selection convey useful information in their own right. In particular, as Grünwald (1999, pp. 94-95) explains carefully, the value of w^* measures the 'randomness' of the data with respect to the model Y^*, so that smaller values of w^* indicate that the the model provides relatively less information about the data.

Demonstrations of the MDL Criterion

In the remainder of this paper, we present two concrete examples of the MDL criterion evaluating cognitive models, in situations where there is a clear need to assess whether the better goodness-of-fit of some models warrants their additional complexity. The first involves different heuristics for a problem solving task, while the second involves different models of forced-choice decision making.

Optimal Stopping Problem

As a first demonstration of the MDL criterion for deterministic models, consider three different account of human decision-making on an optimal stopping task sometimes known as the full-information secretary problem (see Ferguson 1989 for a historical overview).

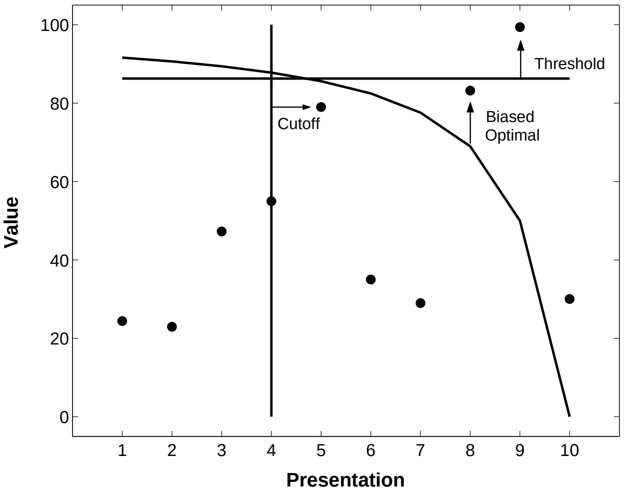

Figure 1: An optimal stopping problem of length 10, with the sequence of values shown by circles, demonstrating the operation of the biased optimal (curved line), threshold (horizontal line) and cutoff (vertical line) models.

Background In these problems, a person presented with a sequence of numerical values, and told to select the maximum. They must decide whether to accept or reject each possibility in turn and, if a possibility is rejected, they cannot select it at some later point. The number of choices in the complete sequence is fixed and known, and the distribution from which the values are drawn (usually a uniform distribution on the interval $[0, 1]$) is also known. Performance is assessed using a 0-1 loss function, so that if choosing the maximum is regarded as correct, but any other choice is regarded as incorrect.

From the mathematical (e.g., Gilbert & Mosteller 1966) and psychological (e.g., Seale & Rappoport 1997) literature, there are at least three plausible accounts of how people might make decisions on these problems. The first 'threshold' model assumes people simply chooses the first value that exceeds a fixed threshold. The second 'biased optimal' model assumes people choose the first value that exceeds a threshold level, where the threshold level changes for each position in the sequence. The threshold levels correspond to the mathematically optimal values (see Gilbert & Mosteller 1966, Tables 7 and 8), for the given problem length, all potentially biased by shifting by the same constant. The third 'cutoff' model assumes people view a fixed proportion of the sequence, remember the maximum value up until this cutoff point, and then choose the first value that exceeds the maximum in the remainder of the sequence. Each of these models has one parameter, giving the threshold, the bias, or the cutoff proportion respectively. For all three models, if no value meets the decision criterion, the last value presented becomes the forced choice.

Figure 1 summarizes the three models on a secretary problem of length 10. The sequence of values presented is shown by the filled circles. The horizontal line shows the constant level used by the threshold model. The threshold levels for the optimal model with no bias follow the solid curve. The vertical line shows the proportion used by the cutoff model. Under these parameterizations, the biased optimal, threshold, and cutoff models choose, respectively, the eighth, ninth, and fifth values presented.

Application of MDL Lee, O'Connor and Welsh (this volume) administered $n = 20$ problems of length $k = 10$ to a number of subjects. For this set of problems, the threshold, biased optimal, and cutoff models are able to predict, respectively, 60, 78, and 9 data sets by varying their parameters. As a concrete example of how the MDL criterion can balance these different model complexities against the fit they are able to achieve, consider the decisions made by one subject from the experiment. For this subject, the best-fitting parameterizations of the threshold, biased optimal, and cutoff models correctly predict, respectively 14, 17, and 10 of the 20 decisions. This is an interesting case to consider, because increases in model complexity lead to increases in model fit.

The MDL criteria values for each model, in relation to this subject's data, are 29.5, 19.4 and 38.0 respectively, showing that, despite its increased complexity, the biased optimal model provides a better account than the threshold and cutoff models. This superiority can be quantified in terms of naturally interpretable odds, because differences between MDL values lie on the log-odds scale. For example, the biased optimal model provides an account that is about $e^{29.5-19.4} \approx 24,000$ times more likely than that provided by the threshold model.

Sequential Sampling Processes

As a second example, we consider the sequential sampling model of decision making developed by Lee and Cummins (in press).

Background Lee and Cummins (in press) proposed that an evidence accumulation approach can unify the 'Take the Best' (TTB: Gigerenzer & Goldstein 1996) model with the 'rational' (RAT) alternative to which it is usually contrasted. The cognitive process being modeled involves choosing between two stimuli on the basis of the cues or features that each does or does not have. In essence, TTB searches the cues until it finds one that only one stimulus has, and then simply chooses that stimulus. The RAT model, in contrast, forms weighted sums across the cues for both stimuli,

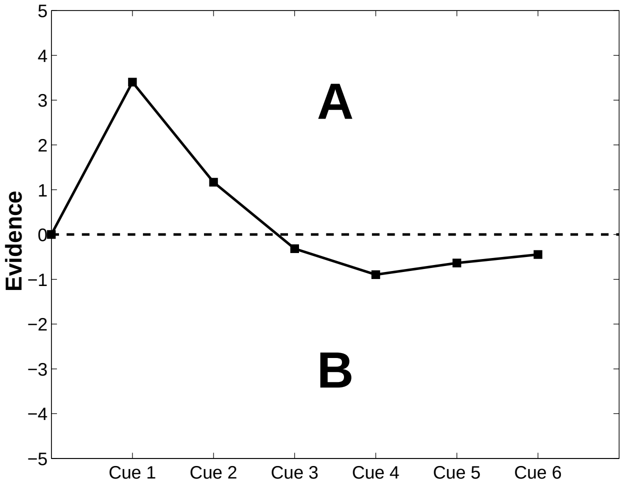

Figure 2: A sequential sampling process using evidence accumulation to decide between choices A and B. Successive evidence values are shown as cues are examined from highest validity to lowest. A decision is made once the evidence exceeds a threshold value.

and chooses the one with the maximum sum.

Figure 2 shows a sequential sampling process accruing information in making this sort of decision. Each of the cues is examined and the evidence provided by that cue is used to update the state of the random walk in favor of choosing stimulus A or stimulus B. If stimulus A has the cue and stimulus B does not, the random walk moves towards choosing A. If stimulus B has the cue and stimulus A does not, the random walk moves towards choosing B. If both stimuli either have or do not have the cue, the state of the random walk is unchanged.

The important observation about Figure 2 is that the TTB and RAT models correspond simply to different required levels of evidence being accrued before a decision is made. If a very small evidence threshold were set, the sequential sampling process would choose stimulus A, in agreement with the TTB choice. Alternatively, if a very large evidence threshold were set, the sequential sampling process would eventually choose stimulus B (because the final evidence is in its favor), in agreement with the RAT model. In general, if a threshold is small enough that the first discriminating cue is guaranteed to have evidence that exceeds the threshold, sequential sampling corresponds to the TTB decision model. If a threshold is large enough that it is guaranteed never to be reached, the final evidence is used to make a forced decision, and sequential sampling corresponds to the RAT decision model.

Application of MDL For the 200 decisions collected from 40 subjects by Lee and Cummins (in press), the TTB model made 36% correctly, while the RAT model made 64% correctly. The sequential sampling model, at the best-fitting value of its evidence threshold parameter, made 84.5% of the decisions correctly. Of course, the sequential sampling model, through its use of the parameter, is more complicated than both the TTB and RAT decision models, which are parameter-free. This raises the issue of whether the extra complexity is warranted by the improved accuracy. Using the model selection method developed here, Lee and Cummins (in press) found MDL values of 87.6, 138.6 and 130.7 for the sequential sampling, TTB and RAT models respectively. The much smaller MDL value for the unified model indicates that it provides a better account of the data, even allowing for its additional complexity.

Conclusion

These demonstration of the MDL criterion provides clear practical examples of how it can be used to evaluate competing deterministic models of human cognitive processes. It also highlights the contribution of this paper, which is a simpler form of the MDL criterion for the special case of 0-1 loss functions. For the optimal stopping problem example, the original MDL formulation involves summing 10^{20} terms in the denominator to find $p\left(D \mid M, Y, w\right)$ for each combination of m and Y that needs to be evaluated in optimization. The simpler form given here requires summing only $n + 1 = 21$ terms each time. For the sequential sampling problem, the original formulation involves $2^{200} \approx 10^{60}$, while the simplification involves 201 terms. As these comparisons make clear, the drastic reduction in computation offered by the simplification developed here makes the MDL evaluation of deterministic cognitive models under 0-1 loss feasible for most (if not) all empirical data collected in cognitive science.

Acknowledgments

I thank Jay Myung, Peter Grünwald and three anonymous reviewers for particularly helpful comments.

References

Ferguson, T. S. (1989). Who solved the secretary problem? *Statistical Science 4*(3), 282–296.

Forsythe, G. E., Malcolm, M. A., & Moler, C. B. (1976). *Computer Methods for Mathematical Computations.* New York: Prentice-Hall.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal Way: Models of bounded rationality. *Psychological Review, 103* (4), 650–669.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *American Statistical Association Journal 61*, 35–73.

Grünwald, P. D. (1999). Viewing all models as 'probabilistic'. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT' 99)*, Santa Cruz. ACM Press.

Grünwald, P. D. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology 44*(1), 133–152.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99* (1), 22–44.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45* (1), 149–166.

Lee, M. D., & Cummins, T. D. R. (in press). Evidence accumulation in decision making: Unifying the 'take the best' and 'rational' models. *Psychonomic Bulletin & Review.*

Lee, M. D., O'Connor, T. A., & Welsh, M. B. (this volume). Human decision-making on the full-information secretary problem. *Proceedings of the 26th Annual Conference of the Cognitive Science Society.*

Luce, R. D. (2000). *Utility of Gains and Losses: Measurement Theoretical and Experimental Approaches.* Mahwah, NJ: Erlbaum.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences 97*, 11170–11175.

Myung, I. J., Forster, M., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology 44*, 1–2.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review 4*(1), 79–95.

Myung, I. J., Pitt, M. A., & Kim, W. J. (in press). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *Handbook of Cognition.* Thousand Oaks, CA: Sage.

Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun., & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, pp. 59–66. Cambridge, MA: MIT Press.

Navarro, D. J., & Lee, M. D. (in press). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin & Review.*

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1990). *The Adaptive Decision Maker.* New York: Cambridge University Press.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory 47*(5), 1712–1717.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103* (4), 734–760.

Seale, D. A., & Rapoport, A. (1997). Sequential decision making with relative ranks: An experimental investigation of the "Secretary Problem". *Organizational Behavior and Human Decision Processes 69*(3), 221–236.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, Similarity, and Bayesian Inference. *Behavioral and Brain Sciences, 24* (4), 629–640.

Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information processing mechanism. *Journal of General Psychology, 124*, 229–339.