A Bayesian Model of Rule Induction in Raven's Progressive Matrices

Daniel R. Little (daniel.little@unimelb.edu.au)
School of Psychological Sciences, The University of Melbourne
Parkville VIC 3010 Australia

Stephan Lewandowsky (stephan.lewandowsky@uwa.edu.au)
School of Psychology, The University of Western Australia
Crawley WA 6009

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley CA 94720-1650 USA

Abstract

Raven's Progressive Matrices (Raven, Raven, & Court, 1998) is one of the most prevalent assays of fluid intelligence; however, most theoretical accounts of Raven's focus on producing models which can generate the correct answer but do not fit human performance data. We provide a computational-level theory which interprets rule induction in Raven's as Bayesian inference. The model computes the posterior probability of each rule in the set of possible rule hypotheses based on whether those rules could have generated the features of the objects in the matrix and the prior probability of each rule. Based on fits to both correct and incorrect response options across both the Standard and Advanced Progressive Matrices, we propose several novel mechanisms that may drive responding to Raven's items.

Keywords: Rule induction, Bayesian inference, Raven's Progressive Matrices

Introduction

Raven's Progressive Matrices (Raven et al., 1998; Raven's from here on) is one of the most widely used assays of fluid intelligence, and much attention has focused on the underlying elemental cognitive processes. Raven's has arguably gathered more attention in the cognitive literature than any other psychometric measure of fluid intelligence, largely because it is an induction task *par excellence* that can be modeled computationally (see e.g., Carpenter, Just, & Shell, 1990; Verguts, De Boeck, & Maris, 2000). For example, Carpenter et al. (1990) presented a production-system model of Raven's to support a two-factor theory of Raven's with working memory capacity (WMC) as the first factor and a second factor related to the ability to abstract relations. This latter ability has been associated with several attributes including rule generation speed (Verguts & De Boeck, 2002), inference speed (Rasmussen & Eliasmith, 2011), and analogical comparison (Lovett, Forbus, & Usher, 2010; McGreggor, Kunda, & Goel, 2010).

These extant models of Raven's have focused on cognitive processes and mechanisms that underlie the inference of rules from the objects in the matrix. Further insight can be gained by exploring a computational-level analysis (Marr, 1982). As performance in Raven's relies primarily on rule induction, the task is conducive to instantiation within a Bayesian framework. For instance, Bayesian models of rule induction have been successfully applied to similar tasks, such as numerical sequence prediction (i.e., which number follows in the sequence: 1, 2, 3, 5, 7, 11?; Austerweil & Griffiths, 2011) and rule-based categorization (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Examining Raven's within the context of a Bayesian model allows exploration of questions about what people's priors (or in non-Bayesian terms, inductive biases) might be like for rules of the variety used in the Raven's test. Finally, the Bayesian formalism provides an extensible framework for using standard extensions to Bayesian models to capture other, more process-based interpretations of factors known to be relevant to performance on Raven's, such as memory and learning.

Here we present a Bayesian model of Raven's which interprets rule induction as Bayesian inference in which a set of rules with some prior probability are evaluated based on their ability to have plausibly generated the features of the items shown in the matrix. Rules are then sampled based on their posterior probability and Bayesian model averaging is used to predict which answers are most likely given the posterior distribution. Unlike extant models, which examine how successful the model is at predicting correct responses (e.g., Carpenter et al., 1990; Lovett et al., 2010; McGreggor et al., 2010), our model also makes predictions about the proportion of responses involving the various incorrect options.

Bayesian Model of Raven's

Solving a Raven's problem can be conceptualized as a three-stage process involving feature extraction, rule-inference and prediction.[1] As illustrated in Figure 1, Raven's items have the following composition:

$$\begin{array}{ccc} O_{11} & O_{12} & O_{13} \\ O_{21} & O_{22} & O_{23} \\ O_{31} & O_{32} & ? \end{array} \tag{1}$$

where O_{ij} is the object in the ith row and j^{th} column. Assuming the features of each object are extracted successfully,

[1] In the present model, we follow Carpenter et al. (1990) by hand-coding the features of the items. Several methods for extracting the features of Raven's items have been proposed (Lovett et al., 2010; McGreggor et al., 2010; Rasmussen & Eliasmith, 2011).

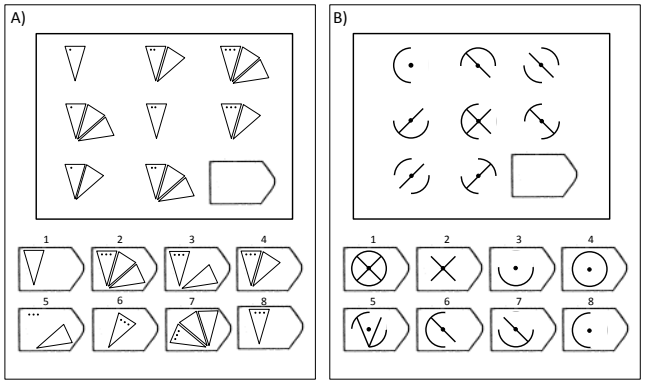

Figure 1: Two examples of matrices like those in the Raven's test. A: Example of an item containing a pairwise incremental rule, a constant rule and a permutation rule. B: Example of an item containing a constant rule and an XOR rule.

such that each object can be decomposed into N features, $O_{ij} = \left\{f_1^{ij}, ..., f_N^{ij}\right\}$, then the goal is to infer the rule that generated the features of the last object in each row and column from the features of the first two objects in each row and column. By design, the rules that are applied to generate O_{13} from O_{11} and O_{12} are the same as the rules used to generate O_{23} from O_{21} and O_{22}. We refer to the set of rules that apply to each row G, where $G = \{g_1, ..., g_M\}$ is a collection of M rules for each feature. The third object in a row is assumed to have been generated by applying these rules to the features of the first two objects: $O_{13} = G(O_{11}, O_{12})$ and $O_{23} = G(O_{21}, O_{22})$. We assume a separate set of rules, H, may apply to each of the features of the objects within a column, $O_{31} = H(O_{11}, O_{21})$ and $O_{32} = H(O_{12}, O_{22})$. The column rules and the row rules may be different; however, because O_{33} can be predicted using either the rows or the columns, we restrict the following analysis to the row rules, G, but it also applies comparably to the column rules, H.

We assume that infering a rule which generated a feature of the third object in a row or column can be conceptualized as finding the posterior probability of each possible rule applied to that feature:

$$p(g|f) = \frac{p(f|g)\,p(g)}{\sum_{i=1}^{M} p(f|g_i)\,p(g_i)} \tag{2}$$

where $p(f|g)$ is the likelihood of generating feature, f, given the rule, g.

Likelihood

In Raven's, all of the rules apply to individual features (which may be discrete or continuous valued; e.g., the three dots in Figure 1, panel A); hence, within a row, the likelihood will have a value of 0 or 1 depending on whether or not the rule successfully produces the features of the third object from the features of the first two objects in that row. To allow for miscalculations in the evaluation of a rule, we set the likelihood equal to ε whenever the rule could not have generated the features of O_{13} and $(1-\varepsilon)$ whenever a rule could have generated the features of O_{13}, where ε is a small number (ε = .01 in our simulations below). We make the further assumption that a rule may work within neither, only one, or both of the rows (or columns), in which case the probability of generating a feature given a rule across both rows is the product of the probabilities from each row separately:

$$p(f|g) = \begin{cases} (1-\varepsilon)^2 & \text{if the rule works for both rows} \\ \varepsilon(1-\varepsilon) & \text{if the rule works for only one row} \\ \varepsilon^2 & \text{if the rule works for neither row} \end{cases} \tag{3}$$

Priors over rules

Carpenter et al.'s (1990) analysis of Raven's identified a taxonomy of rules used to create the Raven's problems. These rule types can be classified as involving transformations (e.g., a quantitative pairwise increment or decrement of a feature from one object in the matrix to the next, or a permutation of objects within a row or column), rules requiring logical operations (e.g., AND conjunctions, OR disjunctions and exclusive-or, XOR, relations between features; Matzen et al., 2010) and a constant rule in which features are maintained unchanged across items.[2] To provide a concrete example, Figure 1 presents two sample Raven's-like problems. The matrix in panel A contains a pairwise incremental rule (i.e., the dots increase across items from left to right) and a permutation rule (i.e,. objects with 1, 2 and 3 triangles are permuted across rows and columns). The matrix in panel B contains a constant rule (i.e., the center dot appears in all items) and an XOR rule (i.e., features which appear in the first two objects do not appear in the third object and features which appear only in one of the first two objects also appear in the third object). Participants must infer these rules from the objects in the matrix and select the missing lower right object in each matrix from the set of possible response options below each matrix.

In total we used eight different rules derived from the taxonomy presented in Carpenter et al. (1990; see also, Matzen et al., 2010) and further analyses: 1) constant, 2) increment or 3) decrement, 4) permutation, 5) logical AND (i.e., maintain common features and delete unique features between objects), 6) logical OR (i.e., maintain unique features between objects), 7) logical XOR (see Figure 1, panel B), and 8) a Distribution of 2 rule. [3]

[2] Carpenter et al. (1990) refers to permutation rules as Distribution of 3 rules because the feature values appear once in the three objects within a row or column and to XOR rules as Distribution of 2 rules because the feature only appears in two of the three objects. Carpenter also refers to logical OR rules as addition and logical AND rules as subtraction.

[3] Six items that were generated using idiosyncratic rules were removed from the 72 Raven's Standard Progressive Matrices (RSPM) and Advanced Progressive Matrices (RAPM) items that we tested. We included the Distribution of 2 rule in our set because there are two items in the RAPM set which use a Distribution of 2 rule that is inconsistent with an XOR rule.

We tested three prior distributions on rules. First, we assumed that each rule had an equal prior probability (i.e., a *uniform* prior probability). Vodegel Matzen, van der Molen, and Dudink (1994) conducted a study using single rule matrices and found there was a clear order of rule difficulty, in which the easiest was constant in a row, followed by quantitative pairwise progression, permutation and logical rule operations, which were the most difficult. To capture this order of difficulty we developed a second prior which assumed that the probability of a rule was proportional to the ease with which that rule could be generated (e.g., the complexity of the rules, which is related to the mental effort necessary to infer and use a rule). For this prior (hereafter referred to as the *Carpenter* prior), we used the frequency with which each rule occured in Carpenter et al.'s (1990) analysis as a proxy for the ease with which each rule could be generated and set the prior probabilities to be proportional to the presentation frequency. Finally, we assumed that the prior may be related to the accuracy with which items containing those rules could be solved. Again, the probabilities are also related to the ease or complexity with which a rule can be generated, but for this prior, we use the relationship between each rule and accuracy on items generated using that rule as a proxy for complexity. To compute this *accuracy-based* prior, we fit a logistic regression model using the rule profile of each item as the predictor variables (i.e., for each item, i, and for each rule, j, we set an indicator equal to 1 if item i was generated using rule j and to 0, otherwise) and the proportion correct for each item as the dependent variable. We then transformed the resulting exponentiated regression weights such that they ranged between 0 and 1 and summed to 1 across all of the rules. The actual probabilities for each of these priors are listed in Table 1.

Table 1: Prior probabilities for each rule.

Rule	Uniform	Carpenter	Accuracy-based
Constant	.125	.194	.150
Increment	.125	.223	.185
Decrement	.125	.223	.141
Permuation	.125	.058	.116
Logical AND	.125	.039	.167
Logical OR	.125	.058	.119
Logical XOR	.125	.165	.119
Distribution of 2	.125	.039	.081

Predicting the response from the posterior

The perceptual complexity of the objects affects how easily rule inferences can be generated (Primi, 2001). For example, Meo, Roberts, and Marucci (2007) showed that performance was significantly worse when features within items were difficult to identify. We incorporate this finding into the current model by assuming that the response is based on the similarity between response options and objects predicted from the rules in the posterior, and for items which have features which are difficult to extract, the similarity does not need to be very high in order for the response options to match the object in the posterior.

Once the posterior probability of each rule is computed for each feature using Equation 2, we compute the missing objects as follows: For a given row rule, G_m, we predict the features of O_{33} by applying the rule to to the features O_{31} and O_{32}. That is, $\hat{O}_{33} = G_m(O_{31}, O_{32})$. Response proportions are determined by computing the relative similarity of each response option, R_k to each object in the predictive posterior by $s_{\{\hat{O}_{33},R_k\}} = \exp\left(-c \times d_{\{\hat{O}_{33},R_k\}}\right)$, where $d_{\{\hat{O}_{33},R_k\}}$ is the Euclidean distance between R_k and $\hat{O}_{33}$ (i.e., the square root of the summed squared differences between the features of $\hat{O}_{33}$ and R_k) and c is a parameter which determines the steepness of the similarity gradient.

Similarities are weighted by their probability in the predictive posterior and normalized across response options to determine the probability that the model chooses each response option (see e.g., Rasmussen & Eliasmith, 2011). In our baseline model, we set c equal to 10 which results in strong responding to the response options which are represented most strongly in the posterior because the similarity gradient is quite steep when $c >> 1$; however, preliminary examination of the model fits to some items suggested that human responses were influenced by the similarity of some distractors to the correct response. For these items, we set c equal to 1, which implies a shallow similarity gradient and greater confusability between similar response options; we exhaustively tested each item to determine whether lowering c improved the fit for that item. We refer to this version of the model as the *Baseline + similarity model*. In this model, 20 of the 66 items had a lower c value than the other items (i.e., for these items, $c = 1$).

Initial inspection of the model predictions additionally revealed a propensity for subjects not to choose response options which also appear as items in the matrix. To handle this, we introduced a heuristic into the model such that all objects that appeared in the matrix were removed from the posterior predictive distribution and from the response set before computing the response proportions. Through exhaustively testing each item, we determined that this heuristic improved the model's predictions for 52 of the 66 items. We refer to this version of the model as the *Baseline + heuristic* model. We additionally tested a *full* model which incorporated both similarity-based responding and the response heuristic.

Table 2: Chi-square values for the fit to choice probabilities for Raven's Standard Progressive Matrices and Advanced Progressive Matrices. The model that provides the best fit to the data for each prior is shown in bold.

Model	Uniform	Carpenter	Accuracy-based
Baseline Model	43603	37363	40699
Baseline + Heuristic	32642	32816	32871
Baseline + Similarity	23385	13884	19948
Full Model	**11761**	**12258**	**9010**

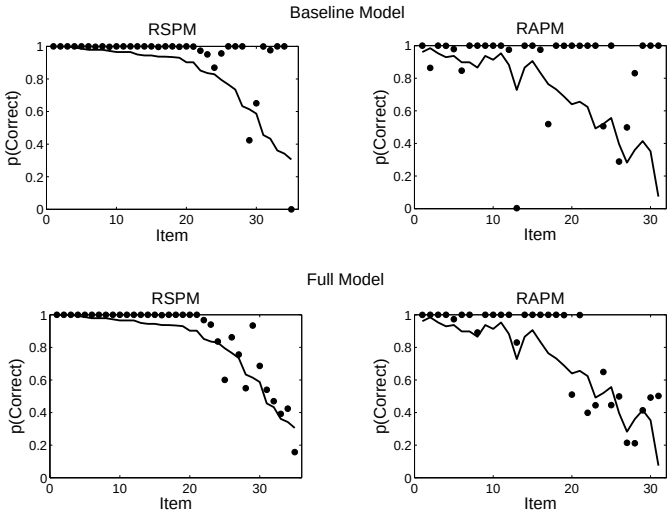

Figure 2: Baseline model and full model fits to proportion correct data from RSPM and RAPM (Little, Lewandowsky & Craig, 2012). Model predictions are shown by the black dots, the observed data are shown by the solid black line.

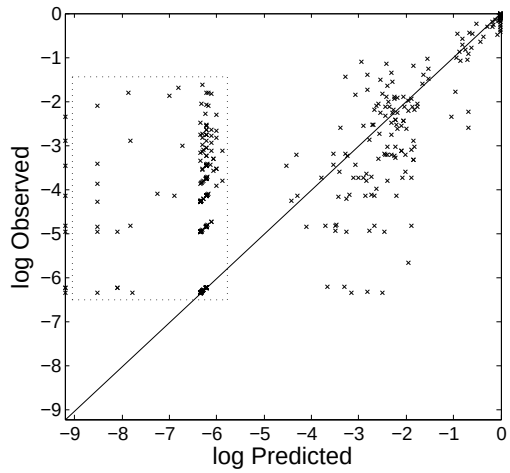

Figure 3: Predicted (full model) and observed response proportions for all response options from RSPM and RAPM. The dotted line surrounds options for which the model predicts near zero response proportions.

Comparing the model to human performance

Descriptive statistics and accuracy data for the Raven's data were previously reported in Little, Lewandowsky, and Craig (2012); here, we are primarily concerned with how often each response option was chosen for each item. (Note that we have removed omissions from this data and look at the distribution of responses across the response options only for participants who actually gave a response). We fit the four versions of the model using the three different prior distributions to the response proportions from the RSPM and RAPM. Chi-square fit statistics are shown in Table 2. Based on these fit values, it is evident that adding similarity-based responding improves the fit over adding the response heuristic to the baseline model. Adding both modifications results in a substantial improvement when a uniform or an accuracy-based prior is used and a marginal improvement when the prior based on Carpenter's analysis is used. The overall best fit is found when the full model is used with an accuracy-based prior; consequently, we now focus on this model's predictions.

Figure 2 shows the accuracy predictions for the baseline and full models for the RSPM items (reordered according to accuracy rate observed in Little et al., 2012) and the RAPM items. The baseline model clearly predicts the correct answer for most of the Raven's items; however, this model overpredicts the propensity with which people choose the correct item for both RSPM and RAPM. By contrast, for the RSPM items, the full model accurately predicts the decrease in accuracy across the items. For the RAPM items, the full model predicts the decrease in accuracy for the hardest Raven's items, but still overpredicts the proportion correct for items in the middle difficulty range.

Figure 3 shows the (log transformed) predictions of the full model (with an accuracy-based prior) across all of the response options against the observed response proportions.[4] For a large proportion of response options, the model predicts the observed proportions correctly; however, the model also erroneously predicts a large number of response options near 0 (predicted log proportion less than -6). These items are also the items for which the model overpredicts the correct response in Figure 2. Examination of the response profiles for individual items reveals that for many items this overprediction is not detrimental to the qualitative pattern of results (see Figure 4). One possible explanation for the discrepant results is that people are guessing the answer with some small probablity which would reduce the accuracy for some of the items and increase the proportion of false alarms to some of the distractors.

Discussion

This paper has defined a Bayesian model of Raven's Progressive Matrices that provides an account of Raven's based on the idea that people infer rules by computing the posterior probability of those rules and using the rules to generate plausible responses. We considered three priors and two ways in which the model could be modified to accomodate human performance. Ultimately, a model incorporating an accuracy-based prior and both modifications provided the best fit to the data.

The success of the accuracy-based prior suggests that rules vary in how they contribute to accurate performance in Raven's. This relationship may reflect sensitivity to differing levels of complexity between the rules, and one way to handle this prior in a more principled way is to instantiate the rules

[4]Proportions equal to 0 or 1 were corrected by setting these proportions equal to $1/(4N)$ or $[N-(1/(4N))]/N$, respectively.

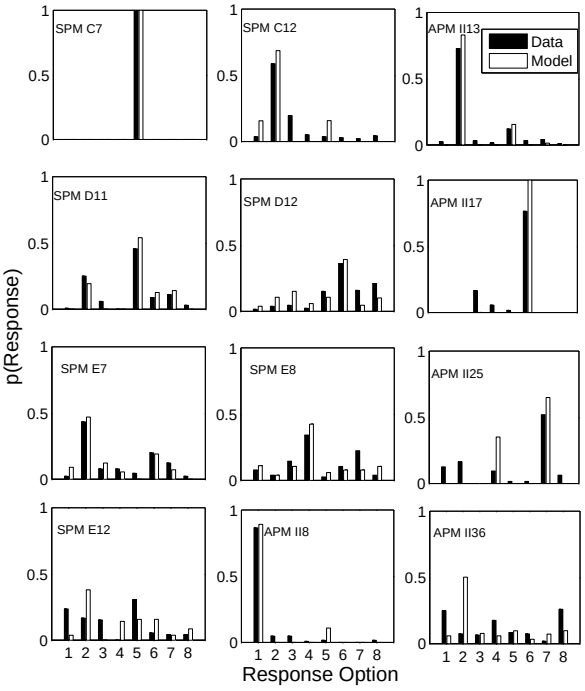

Figure 4: Model prediction profiles for a selection of items. Items SPM C12 (i.e. Standard Progressive Matrices item 12 from Set C) and APM II13 illustrate cases in which the model predicts response probabilities of 0 for responses that people occassionally select. Items SPM E12 and APM II25 illustrate two examples in which the model makes incorrect predictions. For item APM II36, the item with the highest error rate from either test, the model predicts that the correct response (option 2) should be selected most frequently, but humans prefer options 1 and 8.

using a common language with a formal definition of complexity, such as first-order logic (cf. Goodman et al., 2008).

The better fit produced by the similarity-based prediction modification suggests that people vary in how responses are generated to different Raven's items. For instance, for some items, responses are generated by comparing each response option to the possibilities in the predictive posterior; for these items, small differences in features of the response options do not result in a large difference in response prediction. For other items, the response must match the objects in the predictive posterior exactly. An aim for future research would be to identify what makes a feature hard to identify. This would allow appropriate *a priori* specification of the similarity-based prediction rather than the post-hoc approach adopted here. Finally, the heuristic mechanism suggests that people limit their responding to only the plausible response alternatives rejecting alternatives which are implausible because they duplicate items which appear as objects in the matrix.

One of the advantages of formulating a Bayesian model of this task is that we can make use of recent work that has explored how Bayesian models can be extended to capture different aspects of human cognition. One important aspect of performance on Raven's is that differences in WMC correlate highly with accuracy. Furthermore, the correlation with WMC increases as the items become more difficult (if the overall correlation between Raven's and WMC is large enough; Little et al., 2012). The model does not currently incorporate WMC; however, one possibility for extending the model is to represent the hypothesis space as a sampling distribution from the prior using importance sampling.

In an importance sampling scheme (Shi, Griffiths, Feldman, & Sanborn, 2010) samples from the prior are weighted by their likelihood to approximate the posterior distribution; more samples lead to a better approximation of the posterior. Differences in WMC could be modelled by varying the number of samples with high WMC participants having more samples from the prior. With more samples, the model is more likely to generate the correct answer. This idea is reminiscent of the difference between Carpenter et al.'s (1990) FAIRRAVEN and BETTERRAVEN models. The BETTERRAVEN model was given access to more rules than the FAIRRAVEN model and consquently was able to mimic the performance of participants with highly accurate Raven's performance and, by implication, higher WMC.

An alternative account of the relationship between WMC and Raven's is that higher WMC permits faster learning of what rules are likely to be necessary (Verguts & De Boeck, 2002). In support of this account, Carlstedt, Gustafsson, and Ullstadius (2000) found that WMC was correlated more strongly with homogenous intelligence test items (which all required the same rule to solve) than with heterogenous intelligence test items. Presumably, learning the relevant rule is easier for homogenous test items than for heterogenous items. In other related tasks, such as rule-based categorization, WMC is known to be correlated with learning rate (Lewandowsky, 2011; Sewell & Lewandowsky, in press). By this account, the rules at the end of the test are more diagnostic because they have had more time to be learned, thereby leading to greater divergence between low and high ability individuals. Learning in the Bayesian model of Raven's could be instantiated by using a special case of importance sampling known as particle filter sampling (Doucet, Freitas, & Gordon, 2001; Sanborn, Griffiths, & Navarro, 2010). In a particle filter model, a set of particles representing possible rules are drawn in proportion to their prior probabilities. As one progresses through the Raven's items, probabilities are updated in proportion to the success of each rule. Particles representing rules are maintained if they work, but are replaced with new samples from the prior if they do not. Again, higher WMC could be modelled using a larger number of samples. A particle filter model of Raven's would consider both Carpenter et al.'s (1990) and Verguts and De Boeck's (2002) accounts to be correct. That is, higher WMC allows for access to more rules by virtue of allowing more samples from the prior; higher WMC also allows for faster learning by allowing a larger number of particles to be updated from trial to trial.

Consequently, a particle filter model of WMC and Raven's provides a synthesis of these two approaches.

A limitation of Carpenter et al. (1990) and of our own work is that the inputs to the model are hand-coded (Lovett et al., 2010). Hand-coding ignores potentially important spatial representations between objects. Furthermore, Carpenter et al. (1990) did not model the process of rule discovery, but instead fixed the set of rules that were available to the model. The second criticism is less problematic because the rule set is comprehensive, covering the set of rules necessary to handle most of the items; rule discovery is couched in terms of updating the prior probability of each of the rules based on how well those rules work to explain the observed features.[5] Our Bayesian model is susceptible to the first criticism; however, in the present case, we argue that the model provides a good first step toward understanding how the features are used once they are extracted. It is possible that the feature extraction process might be modeled by introducing a prior over features (such as the Indian Buffet process prior, Austerweil & Griffiths, 2010; Griffiths & Ghahramani, 2011). We leave this as a prospect for future development.

Acknowledgments This work was supported by an ARC Discovery Project Grant DP120103888 and an Australian Professorial Fellowship to the second author.

References

Austerweil, J. L., & Griffiths, T. L. (2010). Learning invariant features using the transformed Indian Buffet process. *Advances in Neural Information Processing Systems*, *23*.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*, 499-526.

Carlstedt, B., Gustafsson, J.-E., & Ullstadius, E. (2000). Item squencing effects on the measurement of fluid intelligence. *Intelligence*, *28*, 145-160.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review*, *97*, 404-431.

Doucet, A., Freitas, N. de, & Gordon, N. (2001). *Sequential monte carlo methods in practice*. New York, NY: Springer.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108-154.

Griffiths, T. L., & Ghahramani, Z. (2011). The Indian Buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185-1224.

Lewandowsky, S. (2011). Working memory capacity and categorization: Individual differences and modeling. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 720-738.

Little, D. R., Lewandowsky, S., & Craig, S. (2012). Working memory capacity and fluid abilities: The more difficult the item, the more more is better. *Manuscript submitted for publication*.

Lovett, A., Forbus, K., & Usher, J. (2010). A structure-mapping mode of Raven's Progressive Matrices. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company.

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, *42*, 525-541.

McGreggor, K., Kunda, M., & Goel, A. (2010). A fractal analogy approach to the Raven's test of intelligence. In *AAAI workshops at the 24th AAAI conference on Artificial Intelligence* (p. 69-75). Atlanta, GA: Association for the Advancement of Artificial Intelligence.

Meo, M., Roberts, M. J., & Marucci, F. S. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence*, *35*, 359-368.

Primi, R. (2001). Complexity of geometric inductive reasoning tasks: Contribution to the understanding of fluid intelligence. *Intelligence*, *30*, 41-70.

Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. *Topics in Cognitive Science*, *3*, 140-153.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and vocabulary scales. Section 4: The Advanced Progressive Matrices*. Oxford, UK: Oxford University Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational model: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144-1167.

Sewell, D. K., & Lewandowsky, S. (in press). Attention and working memory capacity: Insights from blocking, highlighting and knowledge restructuring. *Journal of Experimental Psychology: General*.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443-464.

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices. *Journal of Cognitive Psychology*, *14*, 521-547.

Verguts, T., De Boeck, P., & Maris, E. (2000). Generation speed in Raven's Progressive Matrices test. *Intelligence*, *27*, 329-345.

Vodegel Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences*, *16*, 433-445.

[5]We also tested a model with an expanded set of logical rules (e.g., NAND, NOR, etc) but this made no difference to the qualitative pattern of model fits.