Model Flexibility Analysis does not measure the persuasiveness of a fit.

Nathan J. Evans[a], Zachary L. Howard[a], Andrew Heathcote[a,b] and Scott D. Brown[a]

[a] School of Psychology, University of Newcastle, Australia

[b] Division of Psychology, University of Tasmania, Australia

Correspondence concerning this article may be addressed to: Nathan Evans, School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia; Email: nathan.j.evans@uon.edu.au

Abstract

Recently Veksler, Myers, and Gluck (2015, *Psychological Review*) proposed "Model Flexibility Analysis" as a method that "aids model evaluation by providing a metric for gauging the persuasiveness of a given fit". Model flexibility analysis measures the complexity of a model in terms of the proportion of all possible data patterns it can predict. We show that this measure does not provide a reliable way to gauge complexity, which prevents Model Flexibility Analysis from fulfilling either of the two aims outlined by Veksler et al. (2015): absolute and relative model evaluation. We also show that Model Flexibility Analysis can even fail to correctly quantify complexity in the most clear cut case, with nested models. We advocate for the use of well-established techniques with these characteristics, such as Bayes factors, normalized maximum likelihood, or cross-validation, and against the use of model flexibility analysis. In the discussion we explore two issues relevant to the area of model evaluation, being the completeness of current model selection methods, and the philosophical debate of absolute vs. relative model evaluation.

Keywords:

Model Selection; Flexibility; Complexity; Goodness-of-fit

Word Count:

4,859

Introduction

Choosing between competing theories or models is an important aspect of the scientific method. The process of model selection aims to differentiate models based on their predictions, and to select the model that best explains the phenomenon of interest (Myung, 2000; Myung & Pitt, 1997), or that provides the best predictions about future data (Gelman, Hwang, & Vehtari, 2014; Vehtari & Gelman, 2014). This almost always involves balancing two competing qualities of the models. The first is goodness-of-fit, or how well the model predictions match the data. If a model cannot predict the observed data, it is a poor model, but the converse is not always true. For instance, a more complex model may provide a better fit than a less complex model due to its inherently greater flexibility. This highlights the second quality: model complexity, or flexibility. As the principle of "Ockham's razor" points out, a more complex model is a poorer model, all else being equal. The development and use of methods that better account for model complexity is an important and ongoing goal for many fields of science, including psychology.

Recently, Model Flexibility Analysis (MFA) has been proposed (Veksler et al., 2015) as a new way of estimating the complexity of a quantitative model. This estimate of complexity is to be used to answer the question raised in Roberts and Pashler's (2000) seminal paper "How persuasive is a good fit?". MFA examines all the different patterns of data that might be possible, and assesses model complexity by measuring the proportion of those data patterns which can be predicted by the model. For example, if the dependent variables (DVs) of interest in a model are reaction time (RT) and accuracy, MFA involves first listing all possible combinations of RT and accuracy that could be observed, and then asking what proportion of those combinations the model can predict. A complex (or flexible) model, under this framework, is one that can cover a large proportion of the DV space.

The goal of MFA, to highlight the issue of model flexibility by drawing attention to the range of predictions a model can make about the data, is laudable. A model that can predict almost any combination of RT and accuracy is less informative than one that makes a tight prediction about the relationship between those DVs. Quantifying model complexity in this way is important, as some commonly-used model selection techniques do not adequately capture this type of complexity. For example, some methods rely on estimating model complexity by parameter counting (AIC, Burman & Anderson, 2002; BIC, Schwarz et al., 1978), but this can be misleading, because not all parameters are equal.

However, there are important flaws in MFA, including its inability to handle the simple, unambiguous case of "nested" models, its lack of invariance over data transformations, and its unsuitability for a large class of model comparisons. This paper will explore these issues as follows. First, we will give a brief explanation of the theory and application of MFA, and compare it with a conceptually similar approach to model complexity - normalized maximum likelihood (NML). Following that, we will address the most striking failing

of MFA: its inability to successfully handle the nested models problem. Lastly, we will address issues with the two tasks that Veksler et al. (2015) suggested MFA be used for: absolute and relative model evaluation.

In addition to pointing out issues with MFA, we highlight existing, well-established model selection techniques that do not share the problems of MFA, such as Bayes factors (see Kass & Raftery, 1995 for an in-depth explanation) and normalized maximum likelihood (NML; see Myung, Navarro, & Pitt, 2006 or Klauer & Kellen, 2015 for an in-depth explanation). We show that NML, although its shares many similarities with MFA, does not share the flaws we highlight.

Model Flexibility Analysis

MFA aims to assess the complexity of a model through the range of predictions that the model can make. Specifically, MFA involves specifying a model, a range of unique parameter values selected without reference to the data under consideration, and measure(s) (i.e., DVs) of interest. The model is then used to simulate data from every possible combination of these selected unique parameter values. If j is the number of unique values selected for each parameter, these simulations result in j^k joint data predictions for the DVs, where k is the number of parameters in the model. The number the number of cells required to calculate the MFA measure increases exponentially with k, which will present practical difficulties for models with a large number of parameters, but this is a problem common to many model selection methods. The joint data predictions are then used to explore the data space, which is represented as the unit hypercube, $[0, 1]^n$ where n is the number of DVs (i.e., each DV is normalized between 0 and 1). Lastly, a grid is placed across this data space, with $\sqrt[n]{j^k}$ grid cells per dimension, creating j^k grid cells in total. The complexity estimate is then simply the proportion of cells that include at least one model prediction.

To illustrate the MFA procedure, consider applying MFA to models derived from Signal Detection Theory (SDT; see Green & Swets, 1966). Consider an SDT model that has two parameters, d' (sensitivity) and C (criterion), and which produces two key DVs of interest, hit rate (HR) and false alarm rate (FAR). Suppose we select parameter ranges of 0 to 3.25 for d' and -1.6 to +1.65 for C, and intervals of 0.05 for each. This creates 66 unique values (i.e., j) for each parameter, and therefore, 4,356 (66^2) joint parameter pairs, and the same number of data predictions of the model for HR and FAR. To work out the complexity of the model using MFA, we then ask what proportion of the (HR,FAR) combinations in the data space are filled by these predictions. The data space, in this specific example, is a 2-dimensional grid of the possible joint predictions of HR and FAR, with the grid having a height and width of the 66 cells, making each cell 0.015×0.015 units in size. This results in a final complexity value of 0.3985, because the joint data space predictions of the SDT model covers 39.85% (1,736) of the total number of grid spaces (4,356), and therefore, 39.85% of its prediction are unique (and 2,620, or 60.15%,

are redundant) according to the granularity of the given grid.

MFA is not the only procedure that assesses the complexity of a model through the range of predictions that the model can make. For example, normalized maximum likelihood (NML) directly considers all possible data and assesses how well the model fits each of these possible data patterns. This assessment is made via maximum likelihood estimation, which usually entails parameter optimization. In NML, complexity is just the average of these likelihood values, or their integral in the continuous case. To handle very large and small numbers, this is usually represented in log form (Klauer & Kellen, 2015).

Importantly, NML differs from MFA in a few key ways, which point to the cause of the issues with MFA discussed in this paper. Firstly, NML does not require a range of model parameter values to be specified to generate model predicted data from. Instead, NML assesses all possible *data* outcomes, and identifies best-fitting parameters for those. Secondly, NML does not require data summary statistics as DVs, unless those are used in maximum-likelihood estimation (e.g., the SDT model uses HR and FAR to find maximum likelihood parameter estimates). This contrasts with MFA, which uses data summaries in order to reduce the model analysis to a reasonably-sized n dimensional grid. Thirdly, NML requires the model to have a method for maximum likelihood estimation, whereas MFA merely requires a model that can be used to simulate data. This is an advantage for MFA over NML. However, modern methods for approximating model likelihoods can ameliorate this issue (see Discussion section).

MFA can fail to deal appropriately with nested models

If one model can predict all of the data patterns of another model, and some extra data patterns on top, then the larger model is said to nest the smaller. In such simple cases, the nested model is unambiguously less complex than the nesting model. However, even in this unambiguous case, MFA can fail. An example can be seen in the original MFA paper. In their example of applying MFA labeled "Simulation 2", Veksler et al. (2015) compare the MFA complexity values given by three models: the reinforcement learning model, the associative learning model, and the $SAwSu$ model. It is found that the $SAwSu$ model is the least complex of the three, as well as having a superior goodness-of-fit, suggesting that the fit of the $SAwSu$ model is highly persuasive. However, as shown in Veksler, Myers, and Gluck (2014), the $SAwSu$ model is actually a combination of the relative learning and associative learning models, meaning that those models are nested within the $SAwSu$ model. This guarantees that the $SAwSu$ model is more complex than the other two models, because it can predict any data patterns that they can predict, plus others. However, MFA draws the opposite conclusion, and leads the authors to incorrectly overstate the persuasiveness of evidence for the $SAwSu$ model. The $SAwSu$ model must fit at least as well as the other two models, regardless of the data that are observed, because it nests those models. The real question is whether its advantage in fit is more than can be explained by its greater flexibility alone. MFA cannot answer this question, and even

leads the researcher away from the question, by wrongly estimating the persuasiveness of the fit.

The nested-model problem is not unique to the example of the $SAwSu$ model. The problem also arises with signal detection theory. Allowing for a different variance in SDT distributions (known as unequal variance SDT; UV-SDT) produces a model that nests the simpler equal-variance SDT. UV-SDT is clearly more complex as it can accommodate any pattern of data which equal-variance SDT can accommodate, simply by choosing the variances of the signal and noise distributions to be equal. Despite this, as can be seen in Table 1, MFA assigns a *higher* complexity value for equal-variance SDT than it does for UV-SDT (compare rows 1 vs. 8). Again, MFA identifies one model (UV-SDT) as *less complex* than another model (equal variance SDT) that is completely nested within it. By contrast, NML does not have this problem, correctly identifying the UV-SDT model as more complex (Table 1).

Table 1: Eight MFA and two NML analyses of the SDT model (see Supplementary Materials for details). Row 1 shows the MFA analysis of equal-variance SDT, with dependent variables (DVs) hit rate (HR) and false-alarm rate (FAR), and reasonable ranges for the d' and C. Rows 2-8 are MFA analyses of SDT identical to that of row 1, except that: rows 2 and 3 increase the parameter ranges for d' and C, respectively; rows 4 and 5 decrease the parameter ranges; rows 6 and 7 apply exponential and logarithmic transformations of the DVs, respectively. Row 8 analyzes the unequal-variance SDT model, which nests the equal-variance SDT model. Rows 9 and 10 show the NML analysis of the equal-variance and unequal-variance SDT models, respectively. Note that the NML complexity value is calculated by integrating over the log-likelihood of all observable data, meaning that its value can be any real number. However, the principle remains the same, where smaller values indicate less complex models.

	Method	Model	Complexity
1	MFA	Equal variance	0.41
2	MFA	Bigger d' range	0.14
3	MFA	Bigger C range	0.14
4	MFA	Smaller d' range	0.21
5	MFA	Smaller C range	0.17
6	MFA	Exponential transform	0.20
7	MFA	Logarithmic transform	1.0
8	MFA	Unequal variance	0.23
9	NML	Equal variance	-3.9
10	NML	Unequal variances	-3.6

This problem with MFA arises because the number of parameter combinations that must be examined grows when extra parameters are added to a model. As the number of parameter combinations increase, the number of spaces in the MFA grid is increased to

compensate, meaning that models with more parameters will also have a finer grid in the data space, and so more data cells to fill. Figure 1 illustrates this using the example of equal-varaince SDT vs. UV-SDT. This becomes an issue when many parameter combinations fill the same grid cells in DV space, as MFA does not weight each grid cell by the number of predictions it contains, but instead only counts the number of cells containing at least one prediction. Such redundancy (i.e., more than one prediction is a cell) can occur more frequently in the nesting model than the nested model, due to overlapping parameter combinations caused by the increased dimensionality of the nesting model.

Using the equal-variance SDT and UV-SDT example, the number of parameter combinations grows from j^2 in the case of equal-variance SDT to j^3 in UV-SDT. This means that the number of data grid spaces for UV-SDT is j times bigger than for equal-variance SDT, and that for UV-SDT to be considered as complex as equal-variance SDT it must fill j times more grid spaces than SDT. However, due to its additional variance parameter, in the UV-SDT model each combination of the other two parameters, d' and C, which are shared between the models, are repeated j times in UV-SDT (for each value of the variance scaling parameter), resulting in a large number of redundant data predictions that fall within the same grid cells as one another. The final result of this redundancy is that MFA considers the UV-SDT model as less complex, as it fills a smaller proportion of the total grid cells.

A simple potential remedy is to constrain the total size of the data grid to be constant across models. However, this kind of *ad hoc* solution leads to other problems. By enforcing this constraint, the complex model will have a greater number of predicted DV samples than the simple model while being fit into a grid of the same size, meaning the proportion filled will no longer reflect complexity, and instead reflect the number of parameter samples used from the model. Methods to equalize the number of samples from each model – such as randomly taking a number of samples from the complex model's predicted DVs equal to that of the simple model – result in the simple model still being viewed as more complex. Although there may be other potential remedies that can be suggested to this problem, it is clear that the published version is problematic.

Is the measure provided by MFA invariant?

Choosing different DVs to summarize the data, or simulating different, yet plausible, ranges of the parameter values, can have serious consequences for the complexity that MFA estimates. Veksler et al. (2015) propose this as a strength of MFA, as experiments can be found that allow model fits to be persuasive (i.e., have complexity values below some criterion, see their Simulation 4). However, it is also an important weakness of the approach, and shows that MFA frequently varies in its categorization of model's fits as "persuasive" or not.

To illustrate consider again the equal-variance SDT model. Table 1 shows the range of complexity values MFA assigns to this model under different scenarios. Row 1 gives a

complexity measure of 0.41 based on HR and FAR DVs as discussed previously. If we allow a wider range of parameters (rows 2 and 3), MFA estimates a lower complexity value. Such a finding is unambiguously incorrect though, as a wider range of parameter values allows for a greater range of predictions. This is actually another case of nested models which MFA gets wrong – the large-range version of the model nests the small-range version. MFA's inherent inability to factor in redundancy in parameter estimates, as discussed earlier in the section on nested models, creates a scenario in which MFA estimates decreasing complexity when the flexibility of the model unambiguously increases.

The lack of invariance can also be seen in other simple changes. If we decrease the parameter range (rows 4 and 5), MFA again estimates a decrease in complexity value, as the model can now not predict the same range of DV combinations. The changes in MFA-estimated complexity of the SDT model with reasonable changes in parameter range are substantial – "complexity" nearly triples, from 0.14 to 0.41. Additionally, commonly used transformations of the DVs have substantial and unpredictable effects on the MFA-complexity of the SDT model. Table 1 illustrates this with the exponential (row 6) and natural logarithm (row 7) transformations, which approximately halve and double, respectively, the apparent complexity of the SDT model.

Lastly, the choice of data summary variables and the scale on which data values are measured is critical to MFA and the complexity value produced by the method. In many fields, such as those that are interested in the entire distribution of the data, the decision of how to summarize the data into a finite number of DVs has many potential solutions. One such example DV is response time (RT), where different aspects of the distribution of RT are important. A case in point is provided by the EZ diffusion model (Wagenmakers, Van Der Maas, & Grasman, 2007), which is estimated from three data summary statistics: the probability of a correct choice, and the mean and variance of correct response RT. Applying MFA to the EZ diffusion produces very different conclusions based upon whether the variance in RT is included or not (0.02 and 0.28, respectively).

Fortunately, there are well-established methods for model selection that do not suffer from these problems. Bayes factors will only vary with parameter transformations if the prior distributions for the parameters are not subjected to the same transformations. NML separates goodness-of-fit from model complexity, like MFA. However, the transformation of data presents no issue for NML, as this is reflected in a change to the likelihood function; something that MFA ignores. In addition, NML is not vulnerable to the parameter range issues of MFA, as it is based directly in the data space, rather than operating via the parameter space like MFA. Lastly, there is no requirement within the NML framework to summarize the data into specific dependent variables, as is required in the MFA framework, allowing it avoid potential issues associated with the choice of summary variables.

Is MFA appropriate for relative evaluation of models?

The second key part of the MFA approach suggested by Veksler et al. (2015) is that MFA can be used for the relative evaluation of models, where a simpler model providing a superior fit to the data would be highly persuasive, while a more complex model providing a superior fit to the data would be less persuasive. The complexity measure assigned by MFA to a model changes with different transformations of the data, but this might not be problematic for relative evaluations if those transformations changed the complexity of different models equally. That is, MFA's complexity measure could still be useful for model selection if a transformation of the data maintains the *ratio* of complexity for different models, or at the very least, their *order*. Unfortunately, this is not the case.

As an example, we considered the fuzzy logic, linear integration, and signal detection models for perception that have previously been used to test model-selection methods (Myung & Pitt, 1997). We used the same two-by-two design and parameters (one per condition) as Myung and Pitt (1997) (see Supplementary materials for details). When using raw probability as the DV, MFA considers FLMP to be the most complex model, followed by SDT and then LIM (see Table 2, rows 1,2,3). After probit (rows 4,5,6) or logit (rows 7,8,9) transformations of the dependent variables, things that are commonly practiced with probability data, the ratio of complexity between models changes markedly. Even more disturbing, when the data are exponentially transformed (rows 10,11,12), MFA reverses its preference, now considering SDT to be more complex than FLMP, even though the analysis is based on a monotonically-transformed version of exactly the same data. Thus, relative complexity results depend heavily on the data measurement units used in the MFA analysis.

Is MFA useful for Model Selection?

Veksler et al. (2015) provide three simulated examples to show how MFA can be used in combination with a measure of goodness-of-fit to decide upon the best model. Even if MFA's inability to provide an absolute or relative measure of complexity could be remedied we believe its use in model selection is problematic. This is because a quantitative method of combining MFA-complexity with goodness-of-fit is lacking, so it is difficult to apply except in cases where candidate models fit equally well and are only separated by complexity.

Veksler et al. (2015) suggest using MFA in conjunction with a measure of goodness-of-fit – if a simpler model has a higher likelihood then it is highly persuasive, and if a more complex model has a higher likelihood then it is less persuasive. However, it is difficult to see how this could be implemented in an objective way. Exactly how much extra complexity can a better fit justify? This is the kind of question that frequently arises in model selection, and which cannot be addressed by MFA. For example, suppose model A is the best fitting currently available model that falls below the desired level of complexity that would allow its fit to be "persuasive". Further, suppose that model A provides only a relatively poor

Table 2: Complexity values for MFA analyses of the LIM, FLMP, and SDT models from Myung and Pitt (1997). Rows 1-3 show results for probability DVs. All following rows are identical except: rows 4-6 use a probit transformation of the DVs and rows 7-9 a logit transformation (in both cases re-normalised by the maximum value, so to keep all values between 0 and 1, as is required for the MFA analysis). Rows 10-12 report results for exponentially transformed DVs.

	Method	Model	Complexity
1	MFA	LIM	0.1319
2	MFA	FLMP	0.1794
3	MFA	SDT_{percep}	0.1692
4	MFA	LIM Probit transformation	0.2199
5	MFA	FLMP Probit transformation	0.5008
6	MFA	SDT Probit transformation	0.2926
7	MFA	LIM Logit transformation	0.2016
8	MFA	FLMP Logit transformation	0.3907
9	MFA	SDT Logit transformation	0.2273
10	MFA	LIM Exponential transformation	0.0385
11	MFA	FLMP Exponential transformation	0.06
12	MFA	SDT Exponential transformation	0.063

account of the data. Now suppose model B is proposed: it provides an extremely good account of the data, but is slightly above the desired level of complexity. Although model B may justify its extra complexity with a greatly superior goodness-of-fit – as could be assessed by a model selection metric that integrates goodness-of-fit and complexity – the MFA approach might count model B's fit as not being "persuasive", which could lead to the rejection of model B. Although MFA could be used in a situation where the goodness-of-fit of both models is identical, or extremely similar, this situation is quite rare, and it might prove difficult to objectively identify what an acceptable level of similarity would be.

These problems do not afflict other methods of model selection, such as Bayes Factors and NML. For Bayes factors, complexity and goodness of fit are naturally integrated, as the complexity correction is contained within the prior distribution, and therefore influences the marginal probability for each model. For NML, the complexity term is expressed in the same units as the goodness-of-fit term (likelihood), which allows easy combination.

Discussion

One of the main goals of model selection is to choose the best predictive model of a given phenomenon. This requires a metric that accounts for both goodness-of-fit and model flexibility. There are several methods to do this in principled ways. MFA, on the other hand, exhibits a range of flaws. Although Veksler et al. (2015) recommend MFA as

a complement to goodness of fit metrics, rather than a replacement, we contend that these flaws make it unsuitable even for a support role.

Firstly, and most strikingly, MFA is unable to deal with the simplest model complexity issue, nested models. This occurs because MFA fails to account for redundancy in the way parameter combinations produce data patterns. There can be no doubt that a nested model with fewer parameters is less complex than a nesting model with a greater number of parameters, yet MFA can draw the opposite conclusion.

Secondly, MFA assigns varying complexity values depending on the researcher's choices about parameter ranges and data transformations. This makes MFA unable to provide meaningful values for either absolute or relative model evaluation. In addition, MFA is severely limited in scope, since it cannot be combined with goodness-of-fit. We contend that these issues make MFA unsuitable as a tool for model evaluation or selection.

Although MFA is unfit for use as a tool of model selection, some of the underlying principles of MFA are both useful and important. First, the importance of model complexity is often understated, with researchers focusing only on goodness-of-fit, even with complex models. This is problematic, as finding the best model requires accounting for complexity, which is something that Veksler et al. (2015) highlight. Secondly, Veksler et al. (2015) attempted to addresses the key issues of functional-form and data space complexity, which are ignored by some standard metrics (including BIC, AIC, and nested model χ^2 tests).

Fortunately, there are well-established alternative methods that address these issues, yet are not subject to the problems of MFA; two examples discussed in this article are Bayes factors and NML. In most cases, NML is also not much more computationally demanding than MFA. In their standard guises, both NML and Bayes factors require a sometimes difficult-to-compute probability density function for the model, which is a disadvantage relative to MFA – MFA requires only the ability to generate synthetic data from the model. This disadvantage is mitigated by recent work that combines model simulations with a kernel density estimators to approximate a probability density function (Turner & Sederberg, 2012; Turner, Dennis, & Van Zandt, 2013; Turner, Sederberg, & McClelland, 2014; Holmes, 2015). This approach extends the application of methods like Bayes factors and NML to models lacking a tractable probability density function. We recommend that these methods, and not MFA, be used in future research for the crucially important task of model selection taking into account functional form complexity.

Are current methods of model selection complete?

The concepts and issues discussed in this paper also lead to broader, philosophical issues regarding methods of model selection, and model evaluation in general. Even one of the more successful methods previously mentioned within this paper, Bayes factors, is an incomplete metric. Bayes factors ignore an important part of inference: the prior likelihood of each model. When multiplied by the ratio of the prior model likelihoods, the Bayes factor gives the posterior model probability, which may be a better way to select models. This

is important because there are many reasons that scientists hold for preferring one model over another, *a priori*. For example, one model may have greater theoretical plausibility, or more convenient statistical properties, or may have enjoyed a vast amount previous success. These factors might be incorporated into model priors, which might reasonably differ between scientists, for the same model comparison.

These considerations can make it difficult to agree on a reasonable way to identify prior probabilities for different models. A new approach for addressing this issue, and other issues in model selection, uses the concept of data priors. This method takes a conceptually similar approach to MFA, based on a grid over the data space. More specifically, this method involves specifying a prior for the data space, which gives a prior likelihood to each possible data outcome, and then calculating the prior model odd ratio based upon which model best accounts for each potential outcome (by some researcher-defined metric), and the likelihood associated with that outcome.

Data priors allow a variety of different *a priori* model preferences to be incorporated into Bayesian model selection, which confers several advantages. Firstly, this method is able to adjust for the relative ability of the models to account for previous data within this paradigm, by specifying a data prior that is reflective of previous data outcomes and how often they occur. Secondly, this method is able to reflect a researcher's prior preference in a type of model (e.g. simple, useful, most accurate etc.), simply by changing the metric by which the models are judged upon for the potential data outcomes. Based upon these benefits, the method of specifying data priors appears to be a promising avenue for exploration to extend Bayes factors to full Bayesian inference (see Chandramouli & Shiffrin, 2015 for more detail).

Absolute vs. relative evaluation

Another important issue alluded to within this paper is the two broad ways in which models can be evaluated: in an absolute sense, or in a relative sense. Historically, model selection has been largely focused on relative evaluation, with most metrics of model evaluation having little meaning in an absolute sense, and only being meaningful when compared to other models (e.g., AIC, BIC, Bayes factors, etc.). However, relative evaluation has a key shortcoming, that the researcher has no indication of how well each individual model accounts for the data. For example, the selected model could simply be the best of many terrible alternatives, but this will not be apparent from a relative evaluation. This shortcoming indicates a need for absolute methods of evaluation, which focus on how well an individual model can account for empirical trends. The situation is not straightforward though; if there is no alternate model to be evaluated against, then what is the criterion for classing a model as "good" or "bad" at an absolute level?

For example, consider the complexity term for NML. As seen in Table 1, the complexity term given through NML for the equal-variances SDT model is -3.9. However, what meaning does this number have without something to contrast it to? When compared to the NML complexity term of the unequal-variances SDT model, which is -3.6, we can see

that the equal-variances model is the less complex model, showing an ability to make an interpretation in relative evaluation. However, when no point of comparison is present, the complexity term of -3.9 is not useful as an absolute measure of complexity.

One system for absolute evaluation, like that suggested by Veksler et al. (2015) for MFA, would be to use some criterion to evaluate models against, resulting in each model being classed as "good" or "bad" at an absolute level. However, how should that criterion be developed, and how and when it should be changed? Failing to carefully consider these issues properly can have dire consequences for scientific research, similar to the consequences of settling on $\alpha = .05$ for null-Hypothesis significance testing (Leggett, Thomas, Loetscher, & Nicholls, 2013). Unless there is some clear, theoretically based criterion, it appears only sensible to use different criteria in different research areas, based upon how well other models have performed. As an extreme example, models for solar astronomy match the data much more closely than models in cognitive psychology, meaning that using the same absolute evaluation criterion would be unwise. Of course, using a criterion that is based on other models is just another way of moving towards relative model evaluation. Perhaps the primary purpose of absolute evaluation should not be to evaluate whether a model is "good" or "bad", but to discover empirical trends that the best models in the field – found through relative evaluation – cannot capture, and use this as direction for model extensions, such as was done in the many extensions of the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998).

Conclusion

The ability to evaluate models is fundamental to our ability to advance quantitative science. Veksler et al. (2015) proposed a new method, MFA, for evaluating the flexibility and complexity of models, in both an absolute and relative sense. Additionally, Veksler et al. (2015) gave several examples that attempted to show how MFA could be used in selecting between competing models. However, we show that MFA is inappropriate for either absolute or relative model evaluation, due to both its unambigiously incorrect classification of nested models, and its lack of transformation invariance. Additionally, given that MFA has no natural way of integrating with a metric of goodness-of-fit, it will be of little use for model selection even if the above problems were to be fixed. Instead, we recommend that current, well-established methods be used for model selection, such as Bayes factors, NML, or cross-validation.

References

Anderson, N. H. (1981). Information integration theory. *New York*.

Burman, K., & Anderson, D. (2002). *Model selection and multi-model inference: a practical information-theoretic approach*. Springer-Verlag, New York.

Chandramouli, S. H., & Shiffrin, R. M. (2015). Extending bayesian induction. *Journal of Mathematical Psychology*.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Green, D., & Swets, J. (1966). Signal detection theory and psychophysics. 1966. *New York*, *888*, 889.

Holmes, W. R. (2015, December). A practical guide to the Probability Density Approximation improved implementation and error characterization . *Journal of Mathematical Psychology*, *68-69*, 13–24.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Klauer, K. C., & Kellen, D. (2015). The flexibility of models of recognition memory: the case of confidence ratings. *Journal of Mathematical Psychology*, *67*, 8–25.

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. (2013). The life of p:just significant results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2303–2309.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*(2), 167–179.

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

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological review*, *85*(3), 172.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Swets, J. A., & Green, D. M. (1963). *Signal detection by human observers.* (Tech. Rep.). DTIC Document.

Turner, B. M., Dennis, S., & Van Zandt, T. (2013). Likelihood-free bayesian analysis of memory models. *Psychological review*, *120*(3), 667.

Turner, B. M., & Sederberg, P. B. (2012). Approximate bayesian computation with differential evolution. *Journal of Mathematical Psychology*, *56*(5), 375–385.

Turner, B. M., Sederberg, P. B., & McClelland, J. L. (2014). Bayesian analysis of simulation-based models. *Journal of Mathematical Psychology*.

Vehtari, A., & Gelman, A. (2014). Waic and cross-validation in stan. *Submitted. http://www. stat. columbia. edu/˜ gelman/research/unpublished/waic_ stan. pdf Accessed*, *27*(2015), 5.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2014). Sawsu: An integrated model of associative and reinforcement learning. *Cognitive science*, *38*(3), 580–598.

Veksler, V. D., Myers, C. W., & Gluck, K. A. (2015). Model flexibility analysis.

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An ez-diffusion model for response time and accuracy. *Psychonomic bulletin & review*, *14*(1), 3–22.

Supplementary Materials

MFA simulation details

To simulate the standard signal detection theory (SDT; Swets & Green, 1963) model, we allowed two parameters to vary: d' and C. The noise and signal-plus-noise distributions were both assumed to be Gaussian with a standard deviation of 1. The mean of the noise distribution was fixed at 0, with the d' parameter being the mean of the signal-plus-noise distribution, and therefore, the difference in the means of the distributions. The C parameter was specified as the offset from $\frac{d'}{2}$ (i.e. the mid-point between the means of the distributions), so that the absolute criterion was given by the sum of the C parameter and $\frac{d'}{2}$. Specifying the C parameter as an absolute value did not change the pattern of changes when using the MFA approach. To create the parameter space required to create the data space for MFA analysis, d' parameter values varied from 0 to 3.25 in 0.05 intervals, and C parameter values varied from -1.55 to 1.55 in 0.05 intervals, creating a total of 4,158 parameter combinations. For our manipulations of the parameter ranges, the bigger d' range varied from 0 to 20.25 in 0.05 intervals, the bigger C range varied from -1.55 to 10.55 in 0.05 intervals, the smaller d' range varied from 0.5 to 1.25 in 0.05 intervals, and the smaller C range varied from -0.25 to 0.25 in 0.05 intervals.

To simulate unequal variance signal detection theory (UV-SDT), all calculations remained identical, except that we allowed the standard deviation of the signal-plus-noise distribution to vary. The standard deviation of the signal-plus-noise distribution varied in separate intervals when greater than and smaller than the noise distribution's standard deviation, in order to maintain an equal ratio. When greater than the noise distribution's standard deviation, the standard deviation varied from 1 to 3 in intervals of .1, and when smaller than the noise distribution's standard deviation, the inverse of these values were used. With 41 different standard deviation parameter possibilities, there were a total of 170,478 parameter combinations.

The second set of simulations involved three models of perception. The Linear Integration Model (LIM; Anderson, 1981), the Fuzzy Logic Model of Perception (FLMP; Oden & Massaro, 1978), and a perceptual parameterization of the SDT model (Green & Swets, 1966). The functional forms of these models were identical to those used by Myung and Pitt (1997), which were:

$$
\begin{aligned}
LIM: & \\
p_{i,j} &= \frac{\theta_i + \lambda_j}{2} \\
FLMP: & \\
p_{i,j} &= \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1-\theta_i)(1-\lambda_j)} \\
SDT: & \\
p_{i,j} &= \Phi[s_{i,j}\sqrt{|\Phi^{-1}(\theta)^2 + v_{i,j}\Phi^{-1}(\lambda)^2|}]
\end{aligned}
$$

Here, the task of interest was a categorization task with two response options, and two factors, the first with levels $i = 1,..,n$, and the second with levels $j = 1,...,n$. The θ and λ parameters represent the level of support for response option A, given the stimulus information provided by the first and second factors, respectively, both being bounded between 0 and 1. For the SDT model, $\Phi()$ is the standard cumulative normal density function, $s_{i,j}$ is the sign (+/- 1) of $\Phi^{-1}(\theta_i) + \Phi^{-1}(\lambda)$, and $v_{i,j}$ is the sign (+/- 1) of $\Phi^{-1}(\theta_i)\Phi^{-1}(\lambda)$.

For the simulations in our 2x2 design there were 4 dependent variables, being the probability of choosing response option A for each combination of the levels of the two factors. Both θ and λ for each factor level varied between 0 and 0.99, in 11 equally spaced values. Therefore, the total number of parameter combinations was 14,641 (11^4).

Transformations

In order to test the stability of the MFA metric, we subjected the simulated data space to a series of common transformations. Simulated data were transformed after generation; there was no transformation applied to the parameter space. After performing each transformation the data were normalized relative to their maximum, to ensure data remained within the 0-1 range, as required for MFA. Failing to perform such a normalization procedure only resulted in more extreme and variable results. Any data that resulted in an undefined value due to the transformations were removed.

For the logarithmic transformation we took the natural logarithm of the data: $x \mapsto ln(x)$. For the exponential transform we used $x \mapsto e^x$. For the logit transformation: $x \mapsto log\frac{x}{1-x}$. For the probit transformation: $x \mapsto \Phi^{-1}(x)$, where Φ^{-1} is the inverse cumulative normal density function.

MFA calculation

All MFA complexity computations were based on the "*R*" (R Core Team, 2015) code provided by Veksler et al. (2015). We also wrote and tested an equivalent MATLAB version of the code, to ensure the reliability of our findings. All MFA grid sizes were,

again, directly based on the provided code, and thus adhered to the suggested estimation procedure proposed by the authors.

NML simulation details

To assess the complexity metric provided by normalized maximum likelihood, we treated hit-rate and false-alarm-rate as continuous variables, and used the Monte Carlo integration method of Klauer and Kellen (2015), with 500 samples, to estimate the complexity metric, which involves integrating over the maximum likelihood values of all possible data values (i.e., the data space).

To estimate the maximum likelihood that the model could achieve for each sample of the data space, we used the differential evolution algorithm with 300 iterations, 50 particles, and a mutation factor of 0.0001. To improve the search efficiency of the algorithm, the parameters were bounded between reasonable values, being 0 and 5 for d', -2 and 2 for C, and 0 and 5 for the standard deviation of the signal-and-noise distribution, for UV-SDT.

The probability density function was solved using closed-form analytic solutions, with the predicted hit-rate and false-alarm rate being found through the same method as the MFA simulations. From there, the likelihood of the hit-rate and false-alarm-rate given the predicted rates was that of a joint binomial distribution, with no correlation between parameters.

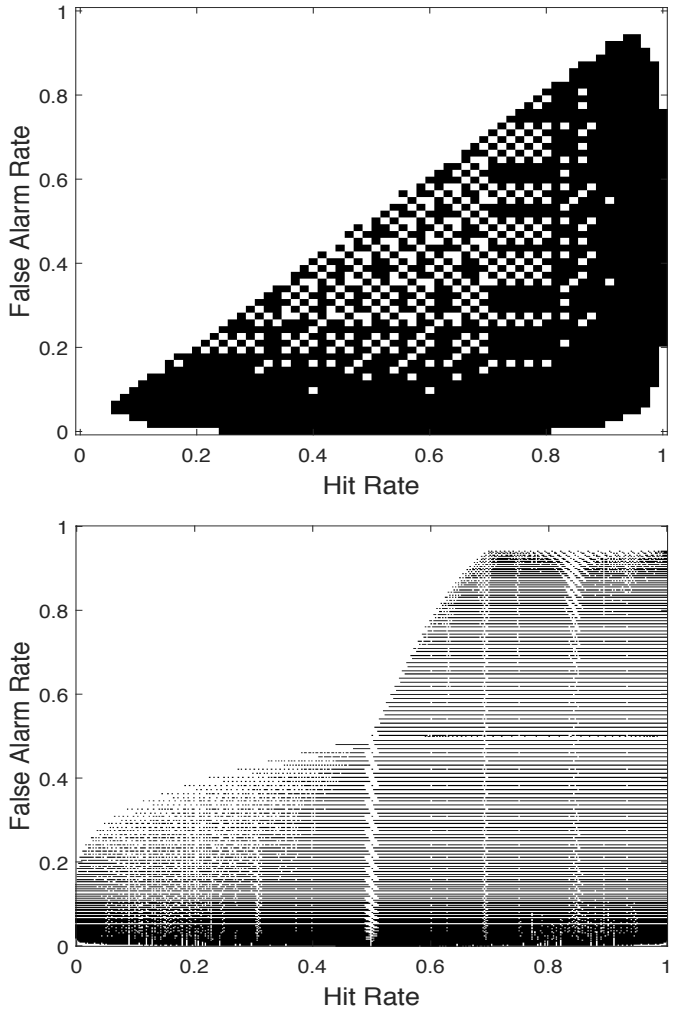

Figure 1. A model flexibility analysis of the equal-variance SDT model (top panel) and unequal variance SDT model (bottom panel). The axes show hit rate (HR) and false alarm rate (FAR), and define the space of all possible data patterns. The SDT model covers 41% of the data space, with the parameters and parameter ranges of that in MFA analysis in Table 1, row 1. The model spans almost half of the grid, with increasingly common "holes" in the grid towards the diagonal, due to an issue with MFA caused by parameter granularity explored further in text. The UV-SDT model uses the parameters and parameter ranges of that in MFA analysis in Table 1, row 9, and clearly covers a much greater portion of this data space than the standard SDT model. However, the model is said to only cover 23% of the data space according to MFA, as the number of "holes" in the grid are substantially greater due to the increased granularity which comes with the greater number of parameter combinations.