Testing for Equivalence: A Methodology for Computational Cognitive Modelling

Terrence C. Stewart  
Centre for Theoretical Neuroscience  
University of Waterloo  
200 University Avenue West  
Waterloo, ON N2L 3G1, Canada  
TCSWART@UWATERLOO.CA

Robert L. West  
Institute of Cognitive Science  
Carleton University  
1125 Colonel By Drive  
Ottawa, ON K1S 5B6, Canada  
ROBERT_WEST@CARLETON.CA

Abstract

The equivalence test (Stewart and West, 2007; Stewart, 2007) is a statistical measure for evaluating the similarity between a model and the system being modelled. It is designed to avoid over-fitting and to generate an easily interpretable summary of the quality of a model. We apply the equivalence test to two tasks: Repeated Binary Choice (Erev et al., 2010) and Dynamic Stocks and Flows (Gonzalez and Dutt, 2007). In the first case, we find a broad range of statistically equivalent models (and win a prediction competition) while identifying particular aspects of the task that are not yet adequately captured. In the second case, we re-evaluate results from the Dynamic Stocks and Flows challenge, demonstrating how our method emphasizes the breadth of coverage of a model and how it can be used for comparing different models. We argue that the explanatory power of models hinges on numerical similarity to empirical data over a broad set of measures.

Keywords: equivalence testing, cognitive modelling, model comparison, dynamic stocks and flows, repeated binary choice

1. Introduction

As scientific theories become more complex, the implications and predictions of a particular theory become increasingly difficult to determine. This has led researchers to rely on computational models to evaluate their theories. These software implementations provide unambiguous descriptions of a theory and allow for quantitative predictions. That is, the outputs of a computer simulation can be compared to empirical results to evaluate the ability of a theory to explain the behaviour of the system of interest.
However, there has been little consensus among researchers as to how to draw conclusions from computational modelling research. This is especially true for models of cognitive systems, where the phenomena to be explained are highly complex, involving many different types of measures and a large degree of statistical variation. This has prompted numerous critiques of computational modelling, including over-reliance on parameter fitting (e.g., Roberts and Pashler, 2000), the difficulties of replicating results (e.g., Axtell et al., 1996), and the use of narrow evaluation criteria (e.g., Simon and Wallach, 1999).

We present a statistical measure and a methodology that addresses these problems by providing a conservative, comprehensive evaluation of the overall similarity between the model and the system being modelled. The remainder of this section discusses what sorts of similarities are important for a model to be taken as a scientific explanation of a real-world system. Section 2 develops a statistical measure to meet these needs. In section 3, we apply this method to two cognitive modelling competitions. From these examples, we present in section 4 a general modelling methodology, showing how this statistic can help identify aspects of a phenomenon that are not yet adequately captured by a model, and thus aid in directing future research.

1.1 The Key Measure for Modelling: Broad Coverage

There are many interacting goals in the scientific process of explaining cognition, and there is a long history of discussion in the philosophy of science as to what counts as evidence. In this section we identify a common thread in much of this literature, the idea that a theory should account for many different kinds of data.

Cognitive modelling is the creation of computational systems whose components are meant to bear some relation to the components of the real system. Bechtel (2006; 2008) identifies these models as mechanistic explanations (or “process models”), where a mechanism is defined as “a structure performing a function in virtue of its component parts, component operations, and their organization” (Bechtel, 2006). Simon and Wallach (1999) argue that due to the fact that they have this complex structure, a good cognitive theory of a given phenomenon will be able to explain many different kinds of empirical findings. This includes not only the basic outcome of an experiment, but also intermediate cognitive steps (e.g., eye movements), reaction times, proportions of different types of errors, effects of changing the context, changes due to learning, and results of physical damage to different parts of the system. All of these aspects can be examined once one has a description of the components and interactions within a system. In contrast, explanations without these internal mechanisms (such as a model based on statistical regression) do not offer any explanation for aspects of the phenomenon other than those for which they were built. Indeed, the explanatory problems that arise from not postulating internal components was a large part of the drive away from behaviourism in cognitive research.

This drive to identify the underlying mechanisms of cognition can also be seen in the work of Pylyshyn (1984; 1989). For him, the goal of cognitive modelling is strong equivalence, where a model not only produces the same observable behaviour as a particular real system, but also performs the task in the same way. To establish this, he identifies reaction times as an important indicator of whether a model is using the same processes as the real system. For example, if one model requires twice as many steps to find an object in a list of five items as in a list of ten items, while another model requires the same amount of steps for both, then the first model implies that the observed reaction time in real people should be greater for a list of ten items than for five items, and the second model implies a constant reaction time for both tasks.
Thagard (1995) argues that the explanatory quality of models can be determined by treating them as analogies and making an inference to the best explanation. If there are broad, systematic correspondences between a model and the real situation (i.e., if it is a good analogy), and if there are no alternative, better analogies available, then one is justified in mapping the explanations from the model to the real system. The best model is one that is able “to explain a lot, to do so simply, to be compatible with what is known, and to be itself explainable” (Thagard, 1995, p. 174). An alternate line of reasoning is provided by Dennett (1991) and expanded on by Ross (2000), where the entities postulated by a model should be treated as real if and only if they are both efficient and predictive. That is, for a model to describe the real components of a system, it must accurately capture observed behaviour, and do so in a compact manner (in an information-theoretic sense). In both of these cases, a model is deemed to be a legitimate explanation (Thagard) or a real entity (Dennett and Ross) if it captures a broad range of the observed data using an efficient, compact representation.

The common theme in all of these arguments is that they require broad coverage of the data by the model. In other words, if a model can accurately predict reaction times, error rates, eye movements, and neural blood flow using the same internal mechanisms, then it will be more convincing than one that can merely predict a few of these.

1.2 Equivalence

Despite the fact that the overall goal of cognitive modelling is to develop models that are equivalent to real-world systems, there are clearly limits on this equivalence. After all, there will always be certain differences between a model and reality. For a mechanistic explanation, the components within the model (and their interactions) are meant to be equivalent to ones in the real system. Importantly, the differences between model and reality in terms of how the underlying components are physically instantiated are not important for this equivalence. A computer model may use transistors and silicon, while the human brain may use neurons and neurotransmitters, but they may still follow an equivalent process, producing equivalent behaviour.

This implies that, when constructing a model of a particular cognitive phenomenon, the choice of basic operations for the various components will place limits on the degree of equivalence of the model. The implementation of each sub-component (such as adding two numbers together, or storing a value) does not need to be done in a manner that is exactly the same as the real system. Certainly, one could create models for these underlying operations, detailing how they might be broken down into even lower-level operations, down to, eventually, neurons (or proteins or atoms or quarks). However, such a breakdown is not usually required. By identifying the level of implementation, the model is assuming that lower level implementation details do not matter for achieving the desired degree of equivalence. If different implementations of these sub-components do not affect the functional or computational aspects of the overall system to a significant degree, then they can safely be ignored.

It is important to recognize that all models will have some degree of error between their behaviour and that of the real system (whether it is due to the choice of operations, the other components, or the algorithm itself). To understand the effect of these discrepancies, it is important to quantify the difference between the model and the real system performance. In particular, if an upper bound can be placed on the size of this modelling error (i.e., if we can be confident that a model is no worse than a certain amount), then this maximum difference can be taken as an indication of the overall quality of a model.
2. Evaluating Equivalence

To satisfy the demands of breadth of coverage (section 1.1) and the desire to quantify the difference between the real system and the model (section 1.2), we develop a new statistical measure called the Maximum Likely Scaled Difference (MLSD). To illustrate the advantages of this approach we compare it to the most common method for model evaluation in cognitive modelling: the Root Mean Squared Difference (RMSD). Similar arguments hold for other typical statistical measures, such as $r^2$ correlation and the $\chi^2$ test (Stewart, 2007).

2.1 Mean Squared Difference

The standard method for evaluating cognitive models is to determine the Root Mean Squared Difference between the model and real-world measures. First, the real-world performance results over multiple trials are averaged together to get a mean value for each different measurement and condition. For example, this could give an average reaction time for each of ten different experimental situations, showing how performance changes in different contexts. Next, the same is done for the computational model; it is run multiple times, adjusting the model to correspond to the changing experimental conditions, and an average is produced for each condition. Finally, these averages are compared to produce an overall average of the difference between the real-world averages and the model averages. This final numerical value is taken as a measure of the quality of the model in that a model with the lowest RMSD is deemed to be the most accurate.

$$RMSD = \sqrt{\frac{\sum (R_i - M_i)^2}{n}}$$

- $R_i$: average real-world performance in condition $i$
- $M_i$: average model performance in condition $i$
- $n$: number of conditions

A fundamental problem with this approach is that it does not take into account the fact that a new source of error has been introduced: sampling error. The model's performance is compared to the performance of the particular people who were the participants in the study and who were each measured a finite number of times. The goal of cognitive research is not to produce a model of those people at that time; rather, the goal is to explain the performance of people in general performing that task. The particular measured average performance on each condition $R_i$ is not the actual average performance of the whole population. If the experiment was repeated a slightly different value for $R_i$ would be found. Any evaluation of the quality of a model needs to address this issue.

To demonstrate the explanatory dangers of not considering sampling error, consider the case of trying to model the flipping of a fair coin. The numerical measure will be '1' for heads and '0' for tails, and the data after 30 coin flips may be as given in Table 1.

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Example data from flipping a coin 30 times. 0 indicates tails and 1 indicates heads.

For the possible models to explain this data, consider a set of simple models, each of the following form:
Output 1 with probability $h$, otherwise output 0.

An infinite number of models can be generated, each with different values of $h$. The correct model is one with an $h$ value of 0.5. However, if these models are each evaluated based on their Root Mean Squared Difference, the model with an $h$ value of 0.6 will have the smallest RMSD. This is because the actual average value of the data provided is 0.6, even though it was produced by flipping a fair coin.

If a large enough sample is made (i.e., if the coin is flipped enough times), then the average will approach 0.5 and then model with an $h$ value of 0.5 will be the closest. However, it is often the case (especially with data from psychological measures) that there will not be an extremely large number of samples. The RMSD approach does not take this into account, and so can provide a false sense of accuracy, focusing attention on one particular “best fit” model when others may, in fact, be more correct. A better measure would be one that could identify which set of models could be correct, and indicate there is insufficient evidence to distinguish between them.

2.2 The Equivalence Test

There is a standard method for statistically comparing two sets of data to determine if their means differ, taking into account the issue of sampling error. This is the $t$-test (and the related ANOVA process). This works by assuming that the two sets of data have the same mean (i.e., the null hypothesis is that $R=M$), and then determining the probability $p$ of finding the amount of difference that was, in fact, observed. If this probability is very small (usually $p<0.05$), then there is said to be a statistically significant difference between the two sets of data.

Unfortunately, this approach is not directly applicable to computational modelling, as it can only be used to detect differences, not similarities. That is, if $p>0.05$, it does not warrant the conclusion that the behaviour of the model matches the behaviour of the real system, as it may be the case that there were not enough samples to reliably detect the difference.

However, the more important problem with using a $t$-test to evaluate models against human data is that a model is always wrong to a certain degree. First, until we achieve a perfect understanding of the human cognitive system, models will always fall short in some way, this is to be expected. Second, in any comparison between model and empirical data, the standard $t$-test or ANOVA will always show a difference if the number of samples from the model is increased sufficiently, unless the model is a perfect model of those particular human subjects performing at that time in that experiment. The difference may be minuscule and irrelevant for practical purposes, but it will still eventually be detected. This problem has led to a number of calls for reducing the reliance on the $t$-test within psychology (Cohen, 1994; Wilkinson et al, 1999; Ioannidis, 2005), including arguments to ban all of this sort of $p<0.05$ testing from psychological journals (Schmidt, 1996).

In short, the $t$-test provides the exact opposite of what is desired. A more useful test for modelling purposes would be one that assumes the model and real data differ by at least a certain amount, and then determines the probability of observing a difference of the size that is, in fact, observed. If this probability is low, then the conclusion could be made that the model is unlikely to differ by more than that certain amount (i.e., it is unlikely that the model is a poor model).

The general problem of evaluating similarity in this manner is sometimes referred to as equivalence testing. Currently, such methods are primarily used in medical testing to determine if a new, cheaper treatment is equivalently effective as an older treatment (Barker et al, 2002). The mathematical derivation is similar to the $t$-test, except that the null hypothesis is that the means of the two groups of data differ by more than a certain amount ($|R-M|>\theta$ rather than $R=M$). If this
assumption leads to a conclusion that the observed data is highly unlikely (usually less than 5%), then the null hypothesis is rejected, leading to the conclusion that the model behaviour and the real behaviour are similar.

A key aspect of this definition is the introduction of a threshold value $\theta$. Differences greater than $\theta$ are deemed to be sufficient for rejecting the model. This allows the researcher to define an acceptable level of error in the model, rather than treating any measurable difference (no matter how small) as grounds for rejection.

The method for calculating an equivalence test can be described in graphical terms, using confidence intervals. A 95% confidence interval is a range around a sampled value such that one can be 95% certain that the actual overall measurement would be somewhere in that range if an infinite number of samples were made. The standard method for calculating confidence intervals assumes that the measures follow a Gaussian distribution. However, we use the bootstrap method (Davison and Hinkley, 1997), as it is not only non-parametric (i.e., makes no assumptions about the distribution of the data), but can also be applied to other statistical measures than the mean. This allows the equivalence test to be applied to any aspect of the statistical measurement, including median, standard deviation, skew, kurtosis, and so on.

To calculate a bootstrap confidence interval for any statistic, the original set of $N$ data points is re-sampled (with replacement) into $B$ sets of $N$ samples each ($B=5,000$ for all results in this paper). For each of these $B$ data sets we compute the desired statistic (for example, the mean). This results in a set of $B$ values which can be ordered from smallest to largest. Removing the top 2.5% and the bottom 2.5% produces the 95% confidence interval.

Given the resulting confidence intervals, the maximum likely difference (MLD; also called the equivalence test threshold) is the maximum difference between the confidence intervals. After all, it is possible that the actual real-world data (if the empirical study were repeated an infinite number of times) is at the high end of its confidence interval, while the actual model data (if the model were run an infinite number of times) could be at the low end of its interval (or vice versa). This identifies the worst-case scenario of how bad the model could be, given the currently available information.

$$MLD = \max (M_U - R_L, R_U - M_L)$$

$R_L$ to $R_U$: real-world 95% confidence interval

$M_L$ to $M_U$: model 95% confidence interval

The resulting MLD is the smallest threshold $\theta$ for which we can reliably reject the null hypothesis that $|R - M| > \theta$. In other words, one can be 95% confident that the model produces behaviour within MLD of the real system, taking into account sampling error. If this value is smaller than the previously chosen threshold $\theta$ for how accurate the model needs to be, then the model and the real system are considered statistically equivalent.

As will be discussed further in section 2.4, the size of the real-world data confidence interval can be seen as a minimum value for $\theta$. There is no sense in searching for models closer to reality that this, since the real-world data is only known to this degree of accuracy. If more accuracy is needed, then more empirical data must be gathered.

Tryon (2001) provides further details on the use of confidence intervals for evaluating differences and equivalence. However, he restricts his discussion to normally distributed data and does not consider the question of combining multiple measures, as is needed to evaluate the breadth of coverage of a model.

The MLD statistic results in a single, interpretable value for comparing the model performance with the subject performance on any aspect of a given measure (e.g., mean, median, standard deviation, etc.). It also provides simple equivalence test: if MLD is less than the desired threshold $\theta$, then the model is statistically equivalent. The MLD value is a worst-case scenario
indication of how bad the model could be: the minimum difference that we can safely assume that there is between the model and reality with 95% confidence. This is not an indication that the model exactly matches the participants’ performance (as no model will ever exactly match to an infinite number of decimal places). Instead, it is an indication that the model is not likely to be in error by more than this amount.

2.3 Multiple Measures

The previous section considered the case where the only basis for comparison is the mean value of a single measure. However, as discussed in section 1.1, an explanatory model needs to perform well on multiple measures across a variety of domains. This not only includes changing the experimental context, but also examining separate measures such as reaction time, accuracy, learning rate, and so on. Furthermore, the distribution of results (standard deviation, median, skew, etc.) should also be considered.

These different measures use different units; reaction time might be in seconds, accuracy might be a percentage, and so on. However, these values can be changed into a common metric. To do this, they can be scaled by dividing each Maximum Likely Difference value by the chosen threshold $\theta$ for that measurement. This results in a value that is less than one if and only if the model is sufficiently equivalent on that measure. If a model has a scaled MLD less than one on all of the measures considered, then it is equivalent on all measures, to whatever thresholds were chosen.

Therefore, if we take a conservative approach and assume that a chain is only as strong as its weakest link, we can judge a model based on the maximum of the scaled MLD values. This measure indicates the worst-case performance. We call the resulting measure the Maximum Likely Scaled Difference or MLSD.

$$MLSD = \max_i \left( \frac{\max(M_{U_i} - R_{U_i}, R_{U_i} - M_{L_i})}{\theta_i} \right)$$

$R_{U_i}$ to $R_{U_i}$: real-world 95% C.I. for measure $i$  
$M_{L_i}$ to $M_{U_i}$: model 95% C.I. for measure $i$  
$\theta_i \geq R_{U_i} - R_{L_i}$: equivalence threshold for $i$

However, there is an important consideration to keep in mind when using multiple measures. Each of the individual measures is only 95% certain (i.e., for each measure, there is a 5% chance that the difference between model and reality is actually larger than the calculated result). If twenty measures are being combined, there is only a 36% chance ($0.95^{20}$) that all of the measures are within the confidence intervals. In other words, even if a model is exactly perfect, there is still a 64% chance that its MLSD will be greater than one over twenty different measures. So one should expect some measures to be inaccurate as the number of measurements goes up.

Given this result, whenever a model has an MLSD greater than one, there is always the possibility that the model is accurate, but the real-world empirical data used for comparison is actually problematic (e.g., strongly affected by outliers). When dealing with multiple measures, this is not only possible, but likely. This makes the task of evaluating a model difficult, since any discrepancy between the model and reality could either be due to the model being wrong or due to the empirical observations being non-representative.

To deal with this, particular measures can be removed from consideration. That is, if it is determined that removing one particular measure produces a model that is equivalent (MLSD<1), then there are two possible interpretations: either the model is not equivalent for that measure, in which case this measure identifies an aspect of the cognitive phenomenon that is not adequately
explained by the model, or that particular measure was a statistical anomaly. In either case, the measures that are removed indicate specific avenues for further research. This ability of the equivalence test to help direct future research will be discussed further in section 4.

2.4 Multiple Models

When comparing different models, each individual model can be considered to be a different hypothesis, \( H \). Running the model and gathering measurements produces the predictions \( P \) from the hypothesis (\( H \rightarrow P \)). If it is determined that \( P \) is false, then it can be concluded that \( H \) is false. However, if \( P \) is true (i.e., if the model is accurately predictive), then one cannot simply conclude that \( H \) is true (i.e., that the model is correct). \( H \rightarrow P \) does not mean that \( \text{not}(H) \rightarrow \text{not}(P) \). Finding one equivalent model does not mean that other models are not going to be equivalent.

In the coin flipping example above, consider only the 11 models with \( h \) values 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1. The 95% bootstrap confidence intervals for the real data and for each of the models (after running each model 100 times) are shown in Figure 1A, and the equivalence test measures (MLD) are shown in Figure 1B.

![Figure 1: Results from a simple model of coin flipping. A) 95% confidence intervals for the mean behaviour of each of 11 models. The shaded area shows the confidence interval of the empirical data from Table 1. B) Results of the equivalence test for the same models. The shaded area indicates the threshold \( \theta \), so models with MLD within this area are statistically acceptable.](image)

In order to draw conclusions from these results, an equivalence threshold \( \theta \) must be chosen. This is a determination of how close a model needs to be to reality for it to be considered sufficiently explanatory. While this value can be chosen by the researcher based on the accuracy desired, there is a minimum size to the threshold that can be used. In particular, the threshold should never be less than the size of the confidence interval of the real-world data. That is, since the available evidence is only sufficient to say that the real average is somewhere between 0.4 and 0.83 (with 95% confidence), then one should not expect models to be provably closer to the real value than 0.83-0.4=0.43. If more accuracy is desired, then more empirical data must be gathered so as to reduce the confidence interval. This process ensures that models do not become over-fit, promising more accuracy than is statistically warranted.

The shaded area in Figure 1A indicates the confidence interval of the coin flips, and the shaded area in 1B indicates the chosen value of \( \theta \). The models with \( h \) values of 0.5, 0.6, and 0.7 are all sufficiently close to the empirical data to be equivalent to this degree; one can conclude with 95% confidence that their results are within 0.43 of the empirical data.
It is crucial to realize that identifying a model as equivalent is a binary measure: a model either is shown to be equivalent or not shown to be equivalent, and all models which are equivalent (MLD<θ) should be treated equally. There is no sense in which the model with an h value of 0.6 is somehow more likely to be explanatory than the other two, even though it does produce a numerically smaller value of MLD. The fact that one particular model has the smallest value is an expected statistical quirk, and not sufficient grounds for a rigorous conclusion. Instead, all three models below the threshold should be treated as equals in terms of their ability to explain the phenomenon, and further distinction between them should rely on introducing new empirical data. It should be noted that the actually correct model (h=0.5) is one of the ones identified to be in the equivalent set, as expected.

2.5 Model Parameters

Thus far, there has been little mention in this paper of the issue of parameters in models. In the previous section, 11 different models were mentioned, each with a different value for h. This sort of situation is common for computational models; numerical values can be adjusted that change the behaviour of the model. Normally, modellers adjust these parameters until a model is found that most closely matches (“best fits”) the empirical results, using a Root Mean Squared Difference measure.

The idea of fitting parameters to particular situations has gathered considerable criticism (e.g., Roberts and Pashler, 2000). Simply searching for the smallest RMSD does not indicate the quality of a model; instead, it may only indicate that a model is flexible enough such that if parameters are adjusted sufficiently, any desired output can be generated. One approach to account for this is to penalize models for having many parameters (e.g., the Bayesian Information Criterion; Schwarz, 1978). This is from the field of Bayesian model comparison, where the intent is to determine whether a set of data supports model A over model B, where each model may have free parameters. In contrast, our goal is to evaluate the similarity between a given model (with fixed parameters) and the real system.

The approach taken here is that if a model we are examining does have free parameters, then we can treat this as a set of different models. Every different set of parameter setting defines a new model with its own behaviour. We can now use MLSD to find the subset of those models which are statistically equivalent to the empirical data. Since these parameters are often continuous values (as in the coin flipping example above), this produces an N-dimensional space of possible models (where N is the number of separate parameters), and regions of this space can be identified as equivalent. Of course, since it is impossible to evaluate every possible parameter setting, this space must be sampled, choosing particular parameter values to evaluate.

Using this approach, a simple, straightforward calculation of the maximum difference between all of the confidence intervals (MLSD) produces a value that captures the overall quality of a model. This allows us to say that a particular model, over a given range of parameters, produces behaviour that is equivalent to the empirical results within a given range and over a given set of measures and statistics. This is exactly the information needed to provide an accurate representation of the capabilities of a model in terms of its breadth of coverage of the phenomenon of interest and a maximum bound on its discrepancies.
3. Applications

To demonstrate the use of the equivalence test, two examples are provided. First, the equivalence test is demonstrated on a model of repeated binary choice. This model subsequently won the Technion Prediction Tournament (Erev et al., 2010) possibly, in part, as a consequence of using this approach to choose the parameter settings. Second, the equivalence test is used to re-evaluate the results of the Dynamic Stocks and Flows model comparison challenge, as described in the introduction to this issue. In this case, we find that since the equivalence test considers worst-case scenario performance (rather than averaging) and a broader range of statistics (rather than just the mean), it produces different results. More generally, we argue that the MLSD provides a useful characterization of the quality of each model, and indicates which aspects of the task are captured well by each model. Other uses of this test can be found in our previous work (Stewart and West, 2007; Stewart, 2007).

3.1 Technion Prediction Tournament

To explore the space of models for decision making tasks, Ido Erev, Eyal Ert, and Alvin Roth organized a prediction competition where computational models could be developed and evaluated based on their ability to predict behavioural results (Erev et al., 2010). Modellers were given a set of empirical results from a binary decision making task (the proportion of time the optimal option was chosen over 60 different experimental conditions), and were free to develop any model of this task. To evaluate these models, results from a separate set of 60 conditions were not provided to the modellers, and the models were evaluated by measuring their Root Mean Squared Difference over this second set of conditions. The competition consisted of three separate tasks and the results discussed here only pertain to the condition where the participants based their choices on repeated experiences with the task.

Although the goal of this competition was merely to be predictive, not to develop a fully explanatory model of this behaviour, we developed a cognitive model for this competition and evaluated it using the MLSD-based equivalence test. Our model used cognitive components defined by ACT-R (Anderson and Lebiere, 1998; Anderson et al., 2004), combined with a blending-based declarative memory system (Gonzalez et al., 2003) and sequential dependencies (Lebiere and West, 1999; see Stewart, West, and Lebiere, 2009 for further details). The model had two continuous parameters affecting the behaviour of the system: the amount of random noise affecting stored memories and the activation threshold for a memory to be recalled. All other parameters for varying the model were fixed by the choice of the ACT-R architecture, or were found to not have any measurable effect on the model performance.

Adjusting these two parameters produces a large space of models that must be sampled and evaluated to find a set of models that produces results equivalent to the provided empirical data. For each of these models, MLSD is calculated, and values below 1 are deemed equivalent. Importantly, since there are sixty experimental conditions being considered, it is expected that even a perfect model would provide MLSD values above 1 for three of the conditions, due to sampling error (more than six errors, however, would be rare, happening less than 3% of the time for a perfect model). As discussed previously, a measure above 1 can be due to a problem in the data or because there is a particular aspect of the phenomenon that is not explained by the model.

For this particular case, seven of the sixty conditions had to be removed. For a perfect model, this would be a rare result, so it is reasonable to conclude that some of those seven conditions represent aspects of the phenomenon which the model cannot yet explain. In particular, whenever one of the two choices in a condition was much better than the other, the model's behaviour was not equivalent to the participants'. This indicates that perhaps an alternate strategy...
or cognitive system is employed in such situations. Further research on this model will involve adding such a mechanism.

The resulting MLSD values (after these seven conditions are removed) are shown in Figure 2A. The large white area in the centre of the figure indicates the parameter settings which result in models equivalent to the empirical data. In other words, in order to explain the given data, the parameter values must be somewhere in that area, but there is no way to further distinguish between these models. This is similar to the situation discussed above with the coin-flipping example, where there was insufficient empirical evidence to distinguish between the models with values of $h$ between 0.5 and 0.7.

For comparison, if the Root Mean Squared Difference method is used, the resulting model space is shown in Figure 2B. This indicates that a different area of the parameter space is the “best fit” to the empirical data. The graph shown provides the RMSD with the seven potential outlier measures removed, but it does not change significantly if those are considered, since RMSD averages over all measures.

![Figure 2: A) Maximum Likely Scaled Difference and B) Root Mean Squared Difference for a space of possible models. Each point represents a different model. Darker shading indicates a larger value. The lighter the shading, the better the model. All parameter settings inside the thick line on the left are statistically equivalent to the empirical data, as per the equivalence test. The black dot on the right indicates the best fit model, as per the standard RMSD approach.](image)

Since only one model was allowed to be submitted to the competition, a single set of parameters had to be chosen. The centre of the large white area in Figure 2A was chosen, giving a noise value of 0.35 and a recall threshold of -1.6. Importantly, if the standard RMSD approach was used to select parameter settings, a different model would have been chosen (noise=0.01 and threshold=-1).

The model chosen using this method won the competition. Although it was not chosen by optimizing for predictive accuracy, it was the most predictive model judged by the standard RMSD criteria. After the competition, it was determined that if the RMSD measure had have been used to select the model for submission (i.e. if Figure 2B was used instead of Figure 2A), the resulting model would not have won (Stewart, West, and Lebiere, 2009).

There are two main conclusions to draw from this example. First, using the equivalence test can result in a model that is more predictive than one chosen with RMSD when applied to new domains. This is because the MLSD measure takes into account sampling error, thus avoiding the problem of overfitting. We can think of this as optimizing a model for explanatory coverage, rather than optimizing to fit sample data. Second, scientifically useful conclusions can be drawn
from the equivalence test. Using the test forced us to identify particular measures for which the model is poor. In this case, an alternate decision-making system for identifying situations where one choice is clearly better than the other is needed. Furthermore, rather than identifying a single “best fit” parameter setting, we have identified that adjusting the model's noise parameter from 0.1 to 0.5 and the threshold parameter from -1 to -2 still leaves the model equivalent to human performance. These values are well within the standard ranges used in other ACT-R models. The actual best model (in terms of having the smallest difference between it and reality) is somewhere in that area, but more empirical data would be needed to further narrow it down.

3.2 Dynamic Stocks and Flows

As discussed elsewhere in this issue, a cognitive modelling competition was also run by Christian Lebiere, Cleotilde Gonzalez, and Walter Warwick. As with the prediction competition in the previous section, empirical data was provided on a set of tasks, and modellers were invited to create cognitive models that would be evaluated on a separate set of tasks. In this case, the task involved controlling the flows into and out of a storage tank of water. The goal was to maintain the water in the tank at a particular level. Over 100 time steps, participants chose how much water to let into the tank and how much to let out, while other flows into and out of the tank occurred outside of their control. To be successful, participants had to learn to compensate for these external flows and for any temporal delay between their commands to adjust the flow and when those commands actually took effect.

Figure 3 shows one of the five tasks the models were evaluated on: the Delay=2 condition. In 3A we show the mean participant performance and the mean performances for two of the nine models entered into the competition (models 3 and 5), along with the standard RMSD measurements for each. In 3B we show the alternate equivalence test approach. Here, confidence intervals are calculated, and the MLSD measure identifies the worst-case difference between the participants and the models over the 100 trials.

The intent here is to demonstrate that the RMSD measure does not strongly distinguish between these two models, while the MLSD measure indicates model 5 is much closer to the empirical data than model 3. If we only examine mean differences, as in figure 3A, then both
models are somewhat close to the human performance (thin black line), although neither captures the shape of the graph. However, looking at figure 3B we see that the exact shape of the jitter up and down from trials 40 to 100 is unknown. That is, all we can conclude is that the actual mean human performance is somewhere within that black area, but it may not jitter up and down at all. The wild variation may be due to an outlier in the data set or some real underlying pattern: there is insufficient evidence to distinguish between these cases. However, model 5 is consistently closer to this range than model 3. Model 3 is, at worst, incorrect by a factor of 117.71, while model 5 is, at worst, off by a factor of 4.87.

To rigorously compare these models, we actually want to know whether the difference in RMSD values or the difference in MLSD values is statistically significant. Even though neither RMSD nor MLSD were specifically designed to compare models to each other, we can still perform a standard statistical comparison test between these measures. In this case, we are looking to find a statistically significant difference, and so can use standard techniques. The simplest is to find the confidence interval for the difference between the RMSD values for models 3 and 5 (and the same for MLSD). Since these values were not necessarily expected to be normally distributed, we use the bootstrap confidence interval method described in section 2.2. If the confidence interval for the difference between the measures does not include zero, then there is a statistically significant difference (Efron and Tibshirani, 1993).

For this case, the confidence interval for the difference in RMSD values is -343.5 to 1908.0, while the confidence interval for the difference between MLSD values is 70.1 to 629.7. Thus there is not a statistically significant difference between model 3 and model 5 in terms of RMSD, but there is in terms of MLSD. We believe this shows that MLSD does a better job of capturing the performance of these models on this task than the typical RMSD measure. It should also be noted that the actual DSF competition did not use this exact RMSD measure, but rather a similar one which did not average the results together before calculating the RMSD. This led to values of 25691 (model 3) and 25561 (model 5), which are also very close to each other.

To this point, we have only examined the use of the equivalence test for a single condition, while there were five separate conditions in the competition. It is clear that we can continue to use this approach by finding the maximum of the MLSD values for all five conditions; this indicates how incorrect the model could be, at worst, across all five. However, so far we have only considered the mean performance. If we are truly seeking a cognitive model that explains human performance, we would expect that model to also have the same distribution of performance as humans. This means we should also compare on other statistical measures, such as the standard deviation, median, and skew.

Before examining this data, it should be noted that it is extremely rare for cognitive models to be evaluated on anything other than mean performance. This means that it is unlikely that any of these models were developed to capture these distributional aspects of the empirical data. There was also no indication that the competition would be judged on anything other than standard best-fit criteria, so attempting to account for the variance in human performance could actually reduce one's chances of winning the competition. However, we feel that examining measures such as the standard deviation and the median are important if we want to develop models that explain human performance.

The behaviour of the first, second, and third place models from the competition are shown in Figure 4. For each model, the mean, median, and standard deviation measures are shown for each of the five experimental conditions. The light shaded regions indicate the 95% confidence intervals for the participant performance (i.e., the actual mean, median, and standard deviation of the whole population will only be outside of that range 5% of the time). The darker shaded regions show the same for the model.
The first thing to note is that the MLSD measure indicates that model 2 (which came in third place) produces behaviour closer to that of the participants than either of the first and second place models. The worst model 2 performs is in the Sequence=2n condition, where its median value might differ from the human participants by up to a factor of 81. The other models may be in error by significantly more; the median performance of model 9 in the Delay=3 condition may be incorrect by a factor of 36,309, while model 4 may be incorrect in the same condition by a factor of over a million. It is interesting to note that models 9 and 4 produce vastly greater median performance in that condition than the participants, while producing mean values that are in accordance with the participants. This is a clear demonstration that relying solely on means for evaluating a model may be misleading.

It is also worth noting that model 4 consistently has a standard deviation of zero. That is, its performance is exactly the same every time it is run. This is a fairly common feature for models, and it is (mostly) true of models 7 and 8 as well. Such models are often thought of as a model of the average individual, rather than a model of the population. These models can, of course, be modified to produce more variable behaviour, but in a complex task such as this one, it is unclear what effects adding such variability will produce.
Figure 4: Performance of the three top-placing models in the DSF competition. Light gray shows the 95% bootstrap confidence intervals of the mean, median, and standard deviations of performance over five different conditions. The model performance is shown in dark gray. For each condition and statistical measure, the Maximum Likely Scaled Difference between the human and model values is shown. The overall MLSD (at the top of each section) is the maximum of the fifteen separate MLSD values.

Of the nine models in the competition, only models 2 and 8 produced median results in the \textit{Delay}=3 condition that were somewhat close to the participants' results. All of the other models produced behaviour similar to that shown in Figure 4 for models 9 and 4: an extremely quickly growing line upwards (the exception is model 7, which goes to zero, indicating perfect performance which is much better than the actual participants' performance). These lines may be interpreted as an indication that these models are doing something qualitatively different from what is desired. However, it is also worth remembering that the median is rarely considered when developing models in the first place. Since the competition required modellers to only submit a single model with fixed parameter settings, it is likely that the submitted models were optimized to produce realistic means, while other measures were not considered. Before the competition, participant results for four other conditions were made publically available by the organizers for this sort of optimization purpose. If modellers had used equivalence testing to optimize their model (much as we optimized our winning model in the Technion Prediction Tournament discussed in section 3.1), then it is quite possible that results of the competition would change.

In light of this, we believe that any decision as to whether a model is worth pursing in future research should not be made on the basis of the RMSD and other fitting-to-sample measures. Instead, the MLSD measure can be used to evaluate the overall quality of a model across all tasks. Since the MLSD takes the maximum (rather than average) difference, it forces our attention on the particular aspects of a phenomenon with which the model has the most difficulty. If model parameters can be found that give good performance across all conditions, these values should be preferred over ones that give poor performance on one measure, even if they have a smaller average difference across the measures.

Table 2 presents the MLSD values for each of the nine models over the five conditions and three statistical measures (mean, median, and standard deviation). The top three by this measure are models 2, 8, and 5. Since so many models had difficulty with the median we also show the MLSD considering just the mean and standard deviation, where the winners are 2, 9, and 5.
Table 2: Summary MLSD results from the nine models. Values indicate how incorrect a model may be in the worst case scenario, given the available evidence, with 95% confidence.

<table>
<thead>
<tr>
<th>Competition ranking</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLSD</td>
<td>2.2x10^0</td>
<td>81</td>
<td>172686</td>
<td>1.2x10^6</td>
<td>5911</td>
<td>1.2x10^1</td>
<td>4.3x10^3</td>
<td>400</td>
<td>36309</td>
</tr>
<tr>
<td>MLSD (no median)</td>
<td>2.2x10^0</td>
<td>9.1</td>
<td>339.9</td>
<td>5721</td>
<td>18.3</td>
<td>23656</td>
<td>1.2x10^3</td>
<td>216</td>
<td>12.5</td>
</tr>
</tbody>
</table>

4. Conclusions

The methodology presented here provides a worst-case evaluation of the explanatory capabilities of a model. For the model of repeated binary choice, this resulted in identifying particular conditions for which the model was poor, while finding that on all the other conditions the model’s behaviour was statistically indistinguishable from the participants’ behaviour over a range of parameter setting. Furthermore the model was more predictive of human behaviour in other conditions than a variety of other models which used a more traditional RMSD approach to parameter fitting. For the Dynamic Stocks and Flows task, we showed that since the equivalence test takes into account confidence intervals and statistical measures other than the mean, it can be used to capture a wider variety of aspects of model performance, resulting in an indication that very few of the models in the competition showed a similar median and standard deviation to the empirical data. The one that came the closest to capturing these aspects was model 2, which only placed third using the standard measures.

These examples show how the equivalence test MLSD measure allows for conservative conclusions indicating exactly which aspects of a system a particular model can explain. This provides a new, more rigorous method for evaluating computational models and expressing the scope of their capabilities. Throughout this paper, particular situations were identified in which the researcher could adjust aspects of this methodology. These situations are summarized here to provide guidelines for the use of this methodology.

4.1 Equivalence Modelling Methodology

Initially, a set of measures and a set of models are chosen. These models are evaluated in terms of their equivalence to the empirical data. Once this occurs, a researcher will have a space of models, of which a subset of those models may be identified as producing behaviour that is equivalent to the system of interest. At this stage, there are particular directions in which the research can proceed:

1) The researcher may wish to add more empirical measures.

These measures can be more of the same conditions as were already considered, in which case the new empirical data will either reduce confidence intervals (thus reducing $\theta$), or identify statistical outliers that incorrectly indicated a model was not equivalent on that measure. Alternatively, the added measures can be completely new measures. This will generally have the effect of reducing the set of equivalent models and acts as a more
stringent test, requiring more explanatory power from the models. This is useful when the measures initially chosen by the modeller are simple enough that a wide range of models produce equivalent behaviour.

2) The researcher may remove empirical measures.

This will generally increase the set of equivalent models. The measures that are removed indicate aspects of the phenomenon that are not being explained. Different sets of measures can be removed, allowing for the discovery of models that explain one subset but not another. If groups of measures that are similar in some way are removed, then that similarity identifies a particular aspect of the phenomenon that is not yet explained. In the Technion Prediction Tournament, this led to the discovery that particular conditions (ones where one option was much better than the other) need a new cognitive mechanism to be added.

3) The researcher may elect to add new models.

This can be done in response to finding particular measures that would otherwise need to be removed. Having these measures as a target can help guide the modification of existing models to better account for these aspects of the empirical data. Also, entirely new types of models can also be considered. This is an exploratory process involving the creation of new theories, guided by a particular phenomenon to explain and constrained by having to continue to explain the other empirical measures.

This methodology for evaluating computational models should iterate through these steps, with the models gradually improving. At any given time there is a clear indication of what is and what is not potentially explained, and the set of models that meet that criteria are always known. Conclusions from this approach indicate a set of models can explain a particular set of observed phenomenon to a specified degree of accuracy $\theta$. The result is a precise expression of what aspects of cognition can and cannot currently be explained, and to what degree.

Acknowledgements

We would like to thank Chris Eliasmith for his extensive comments and suggestions. We are also grateful to David Reitter and Hau-yu Wong for answering our various questions about the DSF Challenge.

References


