Equivalence: A Novel Basis for Model Analysis

Terrence C. Stewart (terry@ccmlab.ca)
Robert L. West (robert_west@carleton.ca)
Institute of Cognitive Science, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada

Abstract

As cognitive models are developed that are meant to apply to a broad range of phenomena, it is necessary to evaluate how successfully they do so. This is commonly done by measures such as the Mean Squared Error. We propose and demonstrate an alternate approach based on a measure of statistical equivalence. Instead of using sample means, this method uses confidence intervals, and places an upper bound on how wrong the model may be, given the uncertainties in the data. We apply this to the RELACS model in various different repeated binary choice tasks. We show that the equivalence measure identifies ranges of canonical parameter settings that are equally equivalent. It also identifies experimental conditions that are not yet modelled well.

Keywords: modeling; statistics; equivalence; repeated binary choice; RELACS; parameter spaces; philosophy of modeling

Introduction

When evaluating a cognitive model of a phenomenon, it is common practice to determine how closely the model's behaviour matches that of the real system being modelled. This is typically done through finding the mean squared error, and there is often a determination of a "best fit" parameter setting for the model. Other approaches (e.g., Bayesian methods) requiring assumptions about statistical distributions and prior probabilities are not considered here.

In (Stewart, 2006) and (Stewart, in press), we have argued for supplementing this approach by also determining *the set of equivalent models*. This set consists of all models that *cannot be statistically distinguished* from the original data at a given probability ($p<0.05$). This set of models (one for each possible combination of parameter settings) determines the space of potential *explanations* for the phenomenon at hand.[1] In contrast, the best fitting model (however determined) can be seen as the best *predictor* of the particular data set that was used.

In this paper, we apply this approach to a well-established model of repeated binary choice behaviour in humans. Specifically, we examine the Reinforcement Learning Among Cognitive Strategies (RELACS) model (Erev & Barron, 2005), and determine what potential explanations arise from it. This analysis is contrasted with the more typical fitting analysis performed by the original developers of the model. This work is a part of a project examining various cognitive models of this phenomenon to determine key experiments that may distinguish between them.

Repeated Binary Choice

One of the simplest tasks in experimental psychology is the Repeated Binary Choice (RBC) task. Here, participants make a series of either/or decisions, usually represented by pressing one of two buttons. After each decision, reinforcement feedback of some kind is provided to the participant, indicating how good that choice was. This allows them to refine their choices. For example, choice A may be correct 70% of the time, while choice B is correct the remaining 30% of the time. By giving feedback as to whether the correct choice was made, the subject will eventually choose A more often than B. Of particular interest is how much more often A is chosen over B, and how this changes over time.

In much of the early work on this task (e.g., Myers et al., 1963), the feedback was limited to a simple correct/incorrect indication. In such situations, the overall result in human participants is generally characterized as "probability matching". That is, if A is correct 70% of the time, then the participant will choose A around 70% of the time. This is somewhat surprising, given that the optimal behaviour in this condition is to choose A 100% of the time.

It is also possible to include numerical rewards within the task feedback. This can either involve just a single number indicating the reward given for the button that was pressed (the "minimal information" paradigm), or two values can be provided, indicating both the reward the participant has received as well as the reward they would have received if they had have pressed the other button (the "complete feedback" paradigm). These rewards can be probabilistic; for example, pressing button A may give 1 point half of the time and 10 points the other half of the time, while button B may always give 7 points.

This work has resulted in a large collection of experimental results about the frequency of choices in various conditions. However, the precise pattern of these choices and the mechanism(s) underlying this performance are unclear. Although the idea that participants tend to match probabilities is prevalent, Friedman and Massaro (1998) note that "probability matching in binary choice ... is less robust than most psychologists seem to believe." The actual observed behaviour is much more complex, and a variety of models have been proposed to account for it.

Existing Models

Given the simplicity of this task, a wide variety of existing cognitive models could be applied. Indeed, any reinforcement-based model would be a natural approach, including all Reinforcement Learning (RL) systems.

[1] There are, of course, further criteria that would be needed before a model can be safely considered to be an explanation. Statistical equivalence (to a specified degree) on the set of relevant measures is necessary for an explanation, but is not sufficient.

However, there are a number of robust behaviours observed in Repeated Binary Choice experiments that do not generally occur in RL systems. This has prompted a variety of specific models to account for these effects. Recently, attention has been focused on the RELACS model by Erev and Barron (2005), which was developed to be a descriptive account of this range of results.

The basis of this model is inherent in its name: Reinforcement Learning Among Cognitive Strategies. It consists of three separate learning systems, each of which uses a slightly different approach (Fast Best Reply; Loss Aversion and Case-Based Reasoning; and Diminishing Random Choice and Slow Best Reply with Exploration). A fourth system is in charge of learning which of the three strategies is currently doing the best job (i.e. resulting in the most reward), and on the basis of this deciding which of the three strategies should currently be followed. For more details, see (Erev & Barron, 2005).

It should be noted that the RBC task has also been investigated using ACT-R models (e.g., Fu & Anderson, 2006). As part of our ongoing work, we are examining not only RELACS, but also various ACT-R models, including ones both procedural-memory based and declarative-memory based via sequential dependencies. We are also examining the use of Clarion on this task. For space reasons, we restrict this paper to the RELACS model only.

Measuring Equivalence

The key measure used here to compare model and human performance on this task involves the *equivalence threshold* (E). This indicates how *wrong* the model *could be*, given the available information. This is calculated by examining the maximum difference between the 95% confidence intervals, as shown in Equation 1.

$$E = max(M_U - H_L, H_U - M_L) \qquad \text{(Eq 1)}$$

Here, the model confidence interval is M_L to M_U and the human data's confidence interval is H_L to H_U. As noted in (Tryon, 2001), this is a conservative version of an inverted *t*-test, where H_0:$|\mu_M - \mu_H| > E$ rather than H_0:$\mu_M - \mu_H = 0$ (called an equivalence test in Barker et al, 2002). So, if we use Equation 1 with 95% confidence intervals and find an equivalence of 150 milliseconds, then we can be 95% confident that the model differs from the human performance by no more than 150 milliseconds.

It should be noted that this measure differs considerably from the standard approach of looking at the squared difference between the model's mean performance and the human mean performance. Such a comparison is made in the common Mean Squared Difference (MSD) measure and even in the (less common) use of the correlation between model and human performance. The key limitation with these measures is that they rely on the *sample* means, and are thus a measure of how well the model matches to the particular set of individuals being studied. It must be remembered that the *population* mean (which is what a model should be compared to) is equally likely to be *anywhere* within the confidence interval of the observed data. The equivalence measure is meant to take this into account, and can generalize to any other statistic (such as the variance, median, skew, or kurtosis).

As presented here, the equivalence measure is conservative. In (Tryon, 2001), it is shown that if means are being measured and if we assume the data is normally distributed, these confidence intervals can be reduced by a factor of

$$\text{CI scaling} = \frac{\sqrt{S_M^2 + S_H^2}}{S_M + S_H} \qquad \text{(Eq 2)}$$

However, this scaling factor is not used in our work. The first reason for this is that there is generally much more model data available than real world data (i.e. much higher *N*). This means the scaling factor is generally close to 1 (0.9~0.95), and so has little effect. The main reason, however, is that avoiding this assumption makes the technique applicable to non-normally distributed data, and can thus be applied when comparing statistics other than the mean. This allows us to compare the standard deviations of the model and human performance (or any other measure for which we can determine confidence intervals).

Since the equivalence measure relies on confidence intervals, attaining accurate intervals is important. When raw data is available, we use the bootstrap non-parametric confidence interval (Davison and Hinkley, 1997). This makes no assumptions about the underlying distribution of the data, and is suitable for any statistic. When only summary data is available (as it is for the experimental results discussed herein), we use the standard method for estimating confidence intervals for the mean (using the *t* distribution) and standard deviation (using χ^2).

Multiple Measures

Since repeated binary choice behaviour is revealed only by examining a large number of different experimental conditions, we must have a method for combining data from these conditions. The basic equivalence measure described above is suitable for comparing model performance in *one* experimental condition to that of human participants in that same condition. In order to combine across multiple conditions, there must be some way of determining how well the model performs overall.

In the standard MSD approach, the measures across the different conditions are combined to their mean value. This indicates how close the model is to the real data *on average*. Thus it is possible for the model to be very close on some measures, but much worse than average on others.

Our approach is that, instead of taking the average difference across measures, we use the *maximum* difference. That is, if a model is very close on two measures, but highly different on a third measure, then the overall equivalence for that model should be the value for the third measure. The overall equivalence is thus an indication that *all* measures under consideration are *at worst* different from the human data by the given amount.[2] It is also possible to scale these values before finding the maximum. Scaling by a factor such as the size of the real-world confidence interval ensures that measures with high uncertainty do not dominate the result. This also provides a simple interpretation to the resulting number: if it is below 1, then the model is statistically indistinguishable from the real-world results at the chosen confidence level.[3] The resulting measure is termed relativized equivalence (E_r).

$$E_r = max_i \frac{max(M_{i,U} - H_{i,L}, H_{i,U} - M_{i,L})}{H_{i,U} - H_{i,L}} \qquad \text{(Eq 3)}$$

We are unaware of any other research using this measure for evaluating the quality of a cognitive model. However, a suggestion that this sort of measure might be possible was independently noted in a footnote in (Axtell et al, 1996) for use in model/model comparisons.

Evaluating RELACS

As with most cognitive models, RELACS has a variety of parameters that govern its behaviour (α, β, λ, and κ)[4]. α determines how quickly the exploration strategy learns (larger values are faster). β does the same for the fast reply strategy. λ controls the balance between exploration and exploitation, with larger values indicating less exploration. κ adjusts the loss aversion system, with higher values being more accurate in estimating previous losses.

The different combinations of parameters define a space of different models, each of which may behave differently. In the original work with RELACS, Erev and Barron searched this parameter space and identified one “best fitting” model. This was the model at α=0.00125, β=0.2, λ=8, and κ=4, which had a MSD of 0.0036.

Two things are unclear from this result. First, since MSD is being used, we do not know how accurate the model is on any *particular* measure. There may be a few conditions for which the model gives extremely different results. Indeed, a visual inspection of the plots in (Erev & Barron, 2005, Figures 2-4) reveals that measures 8, 28, and 33 differ by significantly more than the average difference of 0.06. However, none of the original analysis makes any use of confidence intervals, so it is impossible to determine how well the model is performing on these conditions. It may be that a difference this large is merely due to statistical sampling (especially since many of the studies used by Erev and Barron have only 10 to 14 subjects).

The second ambiguity in the original analysis is how well other parameter settings perform. One parameter setting is given as the best numerical match. However, it may be that many other parameter settings perform *just as well*, in a statistical equivalence sense. The fact that one model gives a slightly better match than the others is *not* necessarily an indication of a better parameter setting, as it may be an indication of over-fitting to the particular sample data.

From an equivalence testing perspective, every model with an E_r less than 1 is *equally good* (i.e. they are all within the same confidence level). If the results produce an area of equivalent models within the model space, then the parameter values within that area can be seen as canonical parameter ranges. This is the range of parameter values over which the model behaves similarly to human participants. For a discussion of the concept of canonical parameter values, see (Anderson & Lebiere, 1998).

However, it is also possible to find *disjoint* areas of the parameter space which provide equivalent models. These represent *alternate explanations* of the human behaviour. Once these alternatives are identified, future experiments can be developed to differentiate them.

Also, when dealing with such a large set of measures, it is quite possible that certain parameter settings will result in models that are equivalent on some measures but not others, and vice versa for other parameter settings. In this situation, the model cannot explain *both* measures, but can explain *either*. These occurrences must be identified so that model developers can resolve them (perhaps by adding mechanisms that adjust parameters based on some change between the conditions).

Since it is impossible to evaluate *every* parameter setting, in this work we sample the parameter space using the values shown in Table 1, for a total of 3,456 settings. We have performed explorations outside of this space, but have not found significant changes in behaviour outside these values.

Table 1: RELACS Parameter Values Examined

α	(exploration learning rate) 0 0.01 0.02 0.03 0.04 0.05
β	(quick learning rate) 0 0.01 0.02 0.05 0.1 0.2 0.5 1
λ	(conservativeness) 1 2 4 8 16 32 64 128
κ	(loss aversion) 0 1 2 4 8 16 32 64 128

Human Data

To simplify the comparison between our approach and the standard one, we use the same set of human data as found in (Erev & Barron, 2005). This is a set of 40 different conditions, mostly consisting of previous studies by Erev and his colleagues. Precise details on all these conditions can be found in their paper. The majority of these conditions use the minimal information paradigm (a single numerical reward value shown after each choice), and the rewards are generally particular values with different probabilities. For example, in condition #23, pressing A

[2] The risk of a bad set of data eliminating a good model is handled by selectively removing conditions from consideration.

[3] It is also possible to choose some predefined scaling factor for each measure, indicating *how close* we require the model to be for a particular purpose. This is highly recommended when the empirical data has small confidence intervals.

[4] Other possible changes to RELACS, such as eliminating one of the three strategies, or choosing randomly between them, can be treated as non-numerical parameters, but are not discussed here.

gives a reward of 32 10% of the time and 0 the remaining 90%, while pressing B gives a reward of 3 all the time.

Two conditions (#29 and #30) had to be removed from consideration, since no information was available on their standard deviations, making it impossible to estimate confidence intervals. Also, standard deviation information was only available for the 2nd block of 100 trials in each condition (with the exception of conditions #15 to #20, which give the 4th block), so only this block is considered.

The remaining conditions are divided into four categories, based on the effects being demonstrated. Most of the first 22 conditions (with the exception of #15 to #20) demonstrate variations on the Payoff Variability Effect. This involves adjusting the variation in the outcomes for a given choice, without adjusting the mean outcome. Observed behaviour changes from risk-seeking to risk-aversion depending on whether the variability is associated with the overall best option

Conditions #15 to #20 examine changing the magnitude of the reward. Choice A is correct 60%, 70%, or 80% of the time, and the reward is 1 or 10. This translates into different monetary rewards given to the participants.

Conditions #23 to #25 investigate the under-weighting of rare outcomes. Here, events that are very rare (<10%) seem to not be considered when determining expected outcomes.

The remaining 15 conditions (#26 to #40) deal with the Loss Rate Effect. In these cases, “when the action that maximizes expected value increases the probability of losses, people tend to avoid it” (Erev & Barron, 2005, p. 917). That is, a choice that has a higher expected value in the long run may be chosen less often if it is comprised of many small losses and few large gains.

Results

These four sets of conditions can be examined separately before combining them for an overall understanding of the RELACS model's performance. The goal here is to understand what conditions RELACS can explain, and to see what can be learned about the parameter values.

Since there are 3,456 different parameter settings to be considered, we cannot present all the gathered data about exactly which parameter settings lead to equivalent models. Instead, we present cross-sections of the model parameter space. These cross-sections are created by holding two of the parameters constant and allowing the other two to vary. For consistency, the same parameters are held to the same values to create the cross-sections in each set of graphs. Cross-section (a) is λ=1, κ=0, (b) is α=0.01, β=0.05, (c) is λ=2, κ=32, and (d) is α=0, β=0.1. These values were chosen to maximize the amount of space shown as equivalent to the human data. Of course, a three dimensional display could further improve the data visualization, but there will always be problems with four or more parameters. Showing optimal cross-sections with this methodology allows for identification of interesting areas in any case.

These cross-sections are shown as contour plots of E_r, where darker shading is less equivalence (larger E_r). A black line has been added indicating E_r=1. Every point inside that contour (coloured pure white) indicates a model that gives performance statistically equivalent to the human performance (at $p<0.05$).

Payoff Variability

There were no parameter values that matched for all 16 conditions examining payoff variability. At most 14 conditions are matched; in these cases, #1, #7, #11, and #13 are the most problematic. Conditions #1 and #7 are the simplest cases (A always giving a reward of 11 and B always giving 10), while #11 and #13 are the most complex, giving real-value rewards (with a Gaussian distribution), as opposed to a one of a small fixed number of rewards.

If these four conditions are eliminated from consideration, many parameter settings fit the remaining conditions (72 out of 3,456). Figure 1a) shows that when κ and λ are low (0 and 1, respectively), there are a large group of equivalent models with α at 0.01~0.02, and β anywhere from 0.02 to 0.2. Figure 1b) shows another view of this same cluster, indicating that it extends up to κ of at least 8 and λ≈2. Figure 1d) shows three separate clusters with different values of κ and λ. Figure 1c) shows how the lower-right cluster in 1d) is shaped as α and β vary.

The main result from Figure 1 is that a wide variety of parameter settings produce models that are statistically indistinguishable from the human participants for these conditions. If we note that setting either α or β to zero effectively turns off that component of the RELACS model, we can see that these models function wildly differently, and yet are still overtly similar to the empirical data.

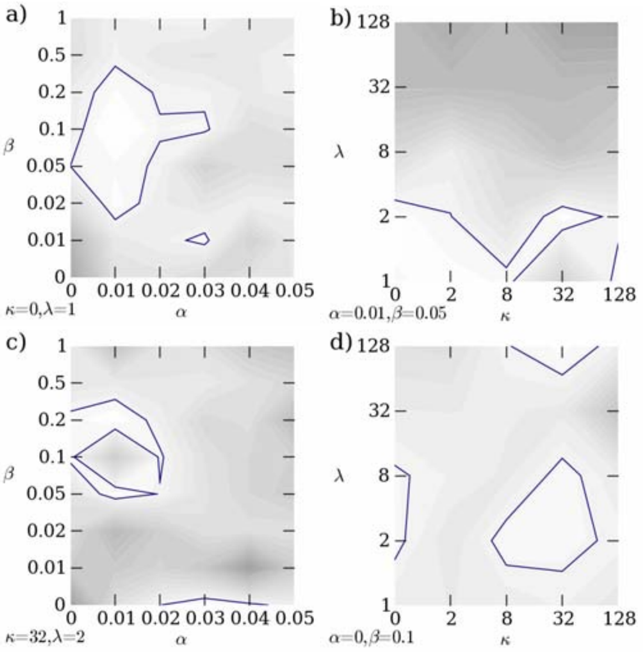

Figure 1: E_r for 10 payoff variability conditions. All points inside the black line (E_r<1) indicate models that are equivalent to human participants, and are shown in white.

Adjusting Reward

The six conditions where reward is adjusted (#15 to #20) are not graphed here. This is because the RELACS model turns out to not change its behaviour *at all* when reward values are scaled. For confirmation, examining (Erev & Barron, Figure 2, p. 915) reveals identical model performance, but changing human data. This fact is not commented on in that paper. It is clear, then, that RELACS cannot account for this aspect of human performance. It should be noted that the MSD approach to analysis used by Erev and Barron did not highlight this fact.

Underweighting

A wide selection of parameter settings are equivalent in terms of the three measures for underweighting. These are shown in Figure 2. Note that 2d) shows that *every* model with α=0 and β=0.1 is equivalent to the human performance, regardless of λ and κ values.

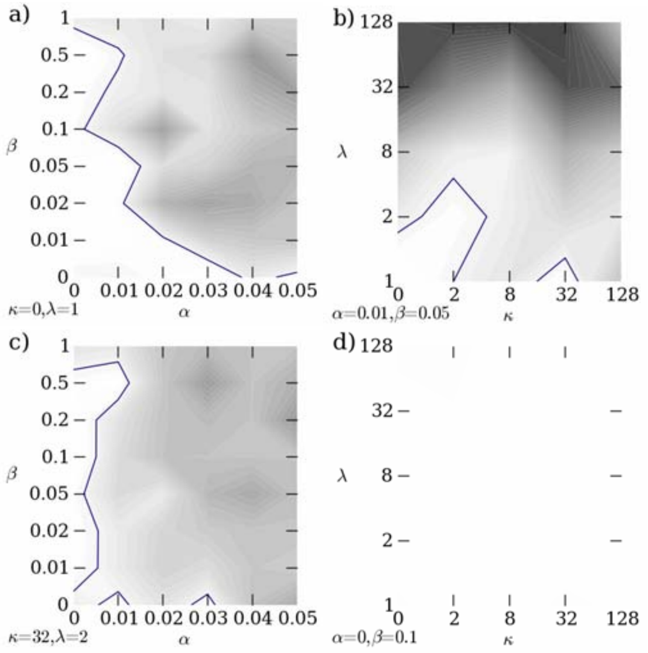

Figure 2: E_r for 3 underweighting conditions. All points inside the black line (E_r<1) indicate models that are equivalent to human participants, and are shown in white.

Loss Rate Effect

As with the payoff variability conditions, there were no parameter settings that were equivalent on all 13 loss rate effect conditions. However, if #28, #33, #34, and #40 are removed, then equivalent models are found for the remaining conditions. These are shown in Figure 3. Note that 3b) indicates only one equivalent model in the very bottom-left of the graph (λ=1, κ=0), and 3c) shows a very few equivalent models (α=0, β=0.1~0.5).

Condition #28 involves a reward with a Gaussian distribution (as did #11 and #13), so the failure to produce good results for this is not surprising given the failure for #11 and #13. Also, #33 and #34 come from a series of studies where the only adjustment is the absolute value of the reward. RELACS does not change behaviour in such situations, explaining its failure here.

However, there seems to be no clear reason why RELACS would fail on condition #40. This measure is a fairly standard experiment where choice A always gives a reward of -3, and choice B gives a reward of -4 80% of the time and 0 the rest of the time. This is merely the opposite of #21, which was modelled well by RELACS.

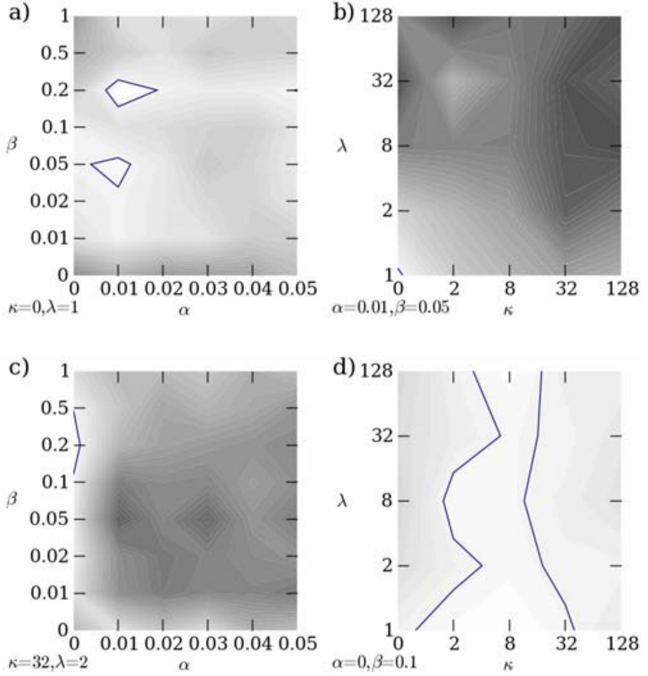

Figure 3: E_r for 13 loss rate effect conditions. All points inside the black line (E_r<1) indicate models that are equivalent to human participants, and are shown in white.

Overall Results

If the measures indicated in the previous figures are combined, we attain Figure 4: an overall plot showing the parameters which give equivalent models on all of the above conditions (except those explicitly eliminated above).

This reveals two small regions of equivalent models. Figures 4a) and 4b) show that the models near (α=0.01, β=0.05, λ=1, κ=0) are indistinguishable from the human data. Figure 4c) and 4d) indicate a separate parameter setting that is also equivalent: (α=0, β=0.2, λ=2, κ=8). Looking at these values, we can see that these models perform very different internal *processes,* yet both result in equally convincing accounts of human performance over this set of RBC task conditions. The first model makes no use of the loss aversion system (κ=0), while the second never adjusts from its initial state in the exploration system (α=0). Both of these models are also *better* (from an equivalence standpoint) than the "best fit" model identified in (Erev & Barron, 2005).

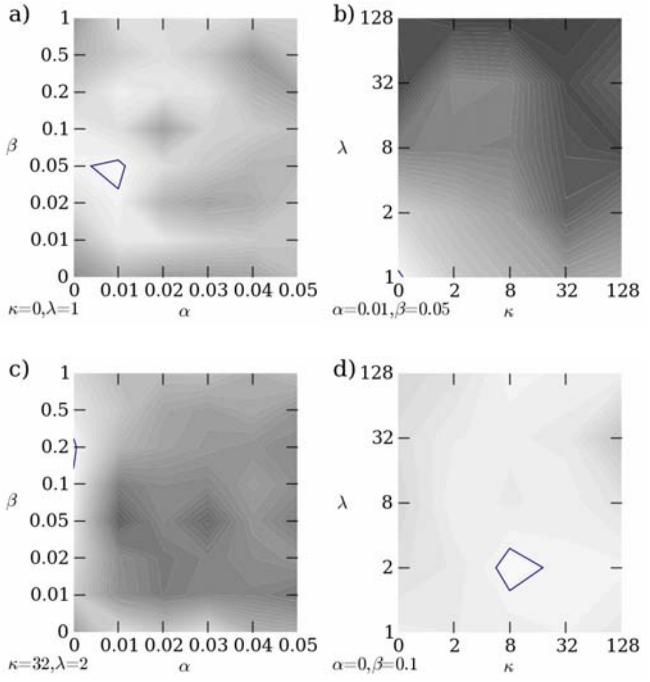

Figure 4: E_r for all 26 remaining conditions. All points inside the black line (E_r<1) indicate models that are equivalent to human participants, and are shown in white.

Discussion

The equivalence method provides us with a collection of models, all of which must be treated equally. These are all parameter settings for models that cannot be distinguished from the real data (at $p<0.05$). These models have differing mean squared error values, but there is no sense in which any of these settings are more equivalent than others (i.e. a better fit may indicate a better model or it may indicate over-fitting the data). This process identifies sets of canonical parameter values that constrain the use of the RELACS model. These parameter values turn out to *not* include the "best " model found by the original researchers.

We have also identified those conditions that are *not* modelled well. RELACS was unable to simultaneously model all of these conditions at $p<0.05$. Instead of averaging across them, we identified those for which RELACS failed. For these, RELACS needs different parameter values for different conditions. Future extensions of RELACS may incorporate mechanisms for detecting these situations and then adjusting its own parameters, but none currently exist.

It is also possible that the failure of RELACS on any of these conditions is due to statistical error, as any of the real-world data sets could be outside the confidence interval (indeed, up to 5% of the conditions may be). Gathering more human data resolves this by exposing atypical results.

We have also restricted ourselves to the *mean* performance only. All of this analysis could also be applied to any other statistic, such as the variance. This has not been done here, as RELACS is known to have a much smaller variance than the human subjects

Conclusion

The equivalence method introduced here supplements the standard "fitting" approach to model evaluation by taking into account confidence intervals and by treating all models that are statistically indistinguishable from the real data as equally good potential *explanations.* We chose the RELACS model to demonstrate this process because the authors had done what few do; they evaluated their model by comparison to a large and diverse set of real-world results. In our opinion, this is critical for evaluating models past a certain level of development. What we have shown in this paper is the value of using an equivalence testing approach for this type of evaluation. Particular conditions can be identified as problematic, and canonical parameter ranges can be identified.

To facilitate further investigation into the RELACS model and other uses of the equivalence approach, all source code, data, and analysis tools used are freely available at <http://ccmlab.ca> as part of the CCMSuite tool-kit.

References

Anderson, J.R. and Lebiere, C. (1998). The Atomic Components of Thought. Mahwah, NJ: Erlbaum.

Axtell, R., Axelrod, R., Epstein, J.R., and Cohen, M.D. (1996). Aligning Simulation Models: A Case of Study and Results. *Computational Mathematical Organization Theory*, 1(2), 123-141.

Barker L.E., Luman E.T., McCauley M.M., Chu Y.R. (2002) Assessing equivalence: An alternative to time use of difference tests for measuring disparities in vaccination coverage. *American J. of Epidemiology*; 156:1056-1061.

Davison, A.C. and Hinkley, D.V. (1997). Bootstrap Methods and Their Application. Cambridge University.

Erev, I. and Barron, G. (2005). On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies. *Psych. Review*, 112(4), 913-931.

Friedman D., and Massaro D. W. (1998) Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, 5, 370–389.

Fu, W-T. & Anderson, J.R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, 135(2), 184-206.

Myers, J. L., Fort, J. G., Katz, L., and Suydam, M. M. (1963). Differential monetary gains and losses and event probability in a two-choice situation. *Journal of Experimental Psychology*, 66, 521–522.

Stewart, T.C. (2006) Tools and Techniques for Quantitative and Predictive Cognitive Science. *28th Annual Meeting of the Cognitive Science Society.*

Stewart, T.C. (in press) Model-Based Science and Artificial Cognitive Systems: The Philosophy of Computational Modelling. *Theoria et Historia Scientiarum: Special Issue on Artificial Life.*

Tryon, W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6(4):371-386.