Nice Graphs, Good R^2, but Still a Poor Fit? How to be more Sure your Model Explains your Data

Niels Taatgen (n.a.taatgen@rug.nl)
Department of Artificial Intelligence, University of Groningen
Nijenborgh 9, 9747 AG Groningen, Netherlands

Hedderik van Rijn (hedderik@van-rijn.org)
Department of Psychology University of Groningen
Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands

Abstract

Although widely criticized, R^2 and RMSE are still the most popular measures to report the quality of fit between model and data. Here we present a different way to assess the quality of fit by comparing the fixed effect estimates of mixed-effects models of both the data and the model. We demonstrate the usefulness of this approach on the basis of a time estimation experiment for which two models were constructed. The model that at first seems to have a superior fit turns out to be based on an invalid characterization of the data when scrutinized more carefully, whereas the alternative model provides an accurate characterization.

Keywords: model fitting; time perception; declarative memory; mixed-effect models

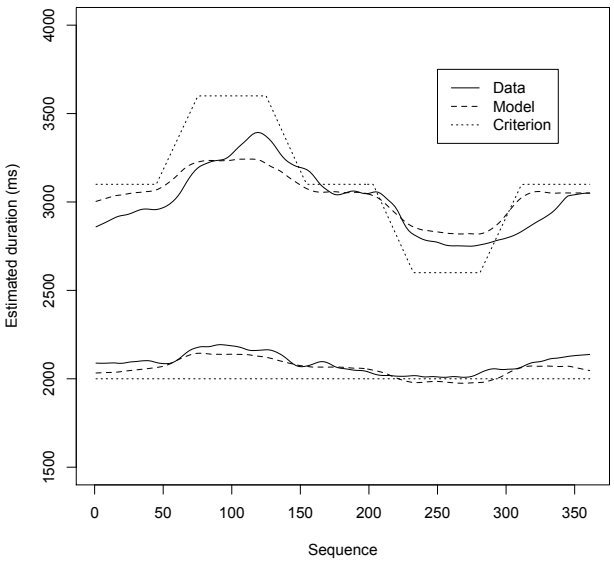
Introduction

One of the unsolved problems in cognitive modeling is how to judge whether a model produces a good fit of the experimental data. Most published papers in which a model is presented try to convince the reader that a fit is good by showing graphs that represent the empirical data along with the model fit. The fit is assumed to be convincing if both graphs are similar. In addition to eyeballing the graphs, statistical measures are often provided to quantify the fit. The most popular measure is R^2, which expresses the correlation between model and data, and some sort of distance measure, like RMSE.

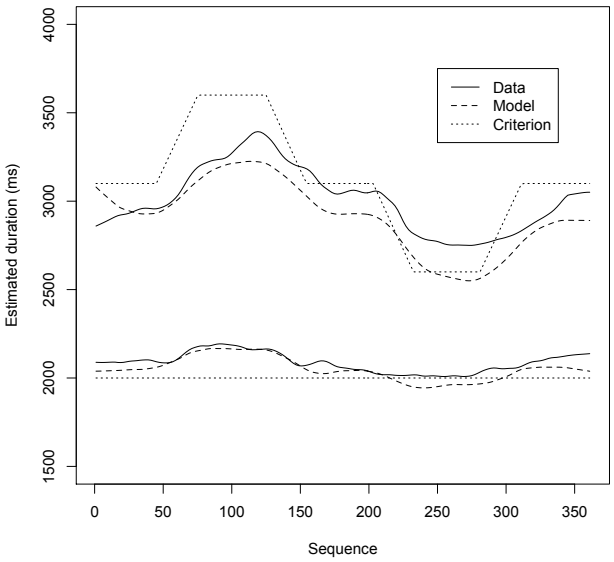
Figure 1 shows an example of two fits between model and data (ignore the "Criterion" curve for now, we will discuss that later). Which of these two models offers a better fit? Neither fit seems to be perfect, but both appear to be reasonable. The following table shows the measures of fit:

Table 1: Measures of fit for the two models in Figure 1

	model A	model B
R^2 upper graph	0.97	0.91
R^2 lower graph	0.81	0.82
RMSE upper graph	178	229
RMSE lower graph	35	46



(a) model A



(b) model B

Figure 1: Two fits between model and data

However, as Schunn and Wallach (2005) pointed out, there is no hard criterion for how high R^2 should be to consider a fit as "good". For RMSE the situation is even less clear, because the measure depends on the measure of the dependent variable. It should just be as low as possible, but there is no standard for what is low enough because the values are dependent on the experiment. Lacking any formal criteria, it is often assumed that the model with higher values for R^2 and lower values for RMSD should be preferred. On the basis of these criteria, Model A should be preferred over Model B, as it outscores it on three of the four measures, and tying it at the fourth. What we will show, though, is that model A is wrong, and model B is reasonably accurate.

Several researchers have criticized the enterprise of fitting models to data. Roberts and Pashler (2000), for example, have pointed out that an ill-constrained model can fit almost any data set. Pitt, Kim, Navarro and Myung (2006) have provided a method to assess the data-fitting capacity of models by examining the partitioning of the space produced by varying all model parameters. If this procedure yields a space with relatively few partitions it means the model makes strong predictions, but if there is a partition for almost any possible outcome, the model is worthless.

Exploration of the parameter space is not always a feasible option, because complex models can take substantial time to run for a single set of parameters, let alone for many combinations. A possible solution to this is to have no free parameters at all, or leave all free parameters at an architectural default, producing so-called "zero-parameter" fits. This is again not always possible, because sometimes parameters have no default value (like some of the parameters in ACT-R's declarative memory), in which case "zero-parameter fits" devolve into "fits with reasonable parameter values". Another issue hidden by the discussion about numerical parameters is the fact that there is considerable freedom in the structural parts of the model (either network topology in neural network models, of symbolic components in a symbolic model). Need another 50 ms to improve the fit? Add a production rule. Need another 200 ms? Add an extra perceptual action. The only way to prevent modelers from wiggling unreported free parameters into their models is to require them to make predictions first and collect data later. The model-data comparison may not always be pretty, but is at least honest (see Taatgen, van Rijn & Anderson, 2007, and Taatgen, Huss, Dickison & Anderson, 2008, for examples).

Apart from the discussion about how a model fit is achieved and how potential alternative fits can be explored, there is the question what kind of measure is a good assessment of a fit. To show that R^2 and RMSD comparisons can deceive, we will first explain our experiment and the goals of the experiment. We will then analyze the data using linear mixed-effect models, and use the same method on the two models. This analysis will provide a better way of comparing models to data, and, although it does not provide absolute criteria, shows convincingly that Model B should be preferred.

Experiment: Memory in Time Perception

To goal of the experiment was to study the role of memory in time perception. In many specialized theories of time perception it is assumed that people are able to represent and store intervals of time in the order of 1 to 60 seconds in memory, without offering any clear theory on the nature of this process. In ACT-R, time perception is modeled using a time estimation module that interacts with the rest of cognition in the same way as other ACT-R modules (Taatgen et al., 2007). The advantage is that ACT-R already has a module for memory, more specifically declarative memory, which can be used to explain memory effects in time perception. We encountered such memory effects in an experiment in which we explored how people estimate partially overlapping time intervals (van Rijn & Taatgen, 2008). In this experiment, subjects had to learn intervals of 2 and 3 seconds, but we noted that the representations of these intervals started to contaminate each other to the extent that some subjects merged both intervals together into a single representation of 2.5 seconds. To study this effect more carefully, we designed a new experiment, of which we will describe one of the conditions here.

Method

In the experiment, subjects learned two intervals, a short one of 2 seconds, and a long one of 3.1 seconds, which they had to reproduce repeatedly, always alternating between the short and the long. Subjects were presented with two circles of the screen, which were gray when they were not active. The circle on the right of the screen was associated with the 2 second interval, while the circle on the left was associated with the 3.1 second interval. During training, one of the circles would change color for a specific duration, and would then turn back to gray. Training consisted of 10 trials, 5 of each duration.

After training, grey circles would again change color to indicate the start of an interval, but now subjects had to press a key to indicate the end of the interval. Subjects received feedback on the accuracy of their produced intervals (we will refer to them as *estimates* from here on): "too short" if they responded earlier than 87.5% of the interval, "too long" if they responded later than 112.5% of the interval, or "correct" otherwise. After training, subjects received 15 warm-up trials of each duration, followed by the experiment proper.

The main manipulation in the experiment is that the criterion for the long interval shifts. For the first 25 estimates of the long interval, the criterion is 3.1 seconds. However, the criterion is then linearly increased to 3.6 seconds over 15 estimates. This means that at some point subjects are told they were too short where they were previously correct. After the shift to 3.6 seconds, the criterion stays at 3.6 seconds, then is decreased back to 3.1 seconds of 15 estimates, stays there for another 25

estimates, then decreases further to 2.6 seconds over 15 trials, stays at 2.6 seconds for 25 trials, increases back to 3.1 seconds over 15 trials and stays there for the remaining 25 estimates. Meanwhile, the criterion for the short interval (remember that short intervals and long intervals are alternated) remains constant at 2 seconds. The "criterion" line in Figure 1 indicates all these shifts. 16 subjects, all students of the University of Groningen, participated in the experiment.

Results

The solid line in Figure 1 shows the mean estimates subjects made for the two intervals. The lines have been smoothed by a Lowess filter (Cleveland, 1981). The results suggest that the two intervals indeed influence each other, given that the changes in criterion for the long interval also impact the estimate of the short interval.

There are (at least) four possible factors that can explain changes in the short interval. One is that the representations of the intervals affect each other directly, i.e., an increase in the internal representation of the longer interval carries over in the internal representation of the short interval. A second explanation is that feedback on the long interval also affects subsequent estimations of the short interval. For example, if we have just produced a long interval, and received the feedback that it was too short, we might unintentionally increase the duration of the short interval that has to be produced next. In addition to the impact of the other interval, previous estimations of the short interval and feedback on those might also impact the next estimate. In order to assess the impact of all these factors, we used mixed-effect models to analyze the data (Baayen, Davidson, & Bates, 2006).

What we did was start out with the most simple regression model to fit the data, and then started adding factors. Each factor adds degrees of freedom to the model, so with each added factor we checked whether improvement in the model was significant with respect to the added degrees of freedom.We started out with the following model, in which the produced short interval is just a constant plus an intercept for each subject:

$$\text{short}_{n,s} = \beta_0 + r_s + \varepsilon_{n,s}$$

So the estimate of short interval n for subject s is equal to constant β_0 plus a random effect for each subject s plus noise. We first start adding the estimates of the previous short intervals. It turns out that including both the previous short interval, and the one before that produce a significant improvement of the model:

$$\text{short}_{n,s} = \beta_0 + \beta_1\text{short}_{n-1,s} + \beta_2\text{short}_{n-2,s} + r_s + \varepsilon_{n,s}$$

Feedback on the previous short estimate also has a significant impact, but not feedback on earlier short estimates:

$$\text{short}_{n,s} = \beta_0 + \beta_1\text{short}_{n-1,s} + \beta_2\text{short}_{n-2,s} + \beta_3\text{short-fb-S}_{n-1,s}\, r_s + \beta_3\text{short-fb-L}_{n-1,s} + r_s + \varepsilon_{n,s}$$

The feedback has two components, because it can be "too short" (short-fb-S) or "too long" (long-fb-L). short-fb-S is equal to 1 if the feedback on the previous trial was "too short", and 0 otherwise. The same is true for long-fb-L and the "too long" feedback. We then added factors associated with the long interval. The estimate of the previous long interval did indeed have a significant impact, but earlier long intervals did not. Finally, we added in the feedback on the earlier long intervals. Here the feedback on the last long interval also led to a significant contribution. Table 2 lists the components and regression values of the final model.

Table 2. Fixed effects in the regression model for the short interval

Fixed Effect	Value of β	*t* value
Intercept	657 ms	4.6
short_{n-1}	0.385	8.3
short_{n-2}	0.085	3.3
short-fb-S_{n-1}	110 ms	3.1
short-fb-L_{n-1}	-208 ms	-6.5
long_{n-1}	0.16	5.1
long-fb-S_{n-1}	92.6 ms	3.2
long-fb-L_{n-1}	-163 ms	-4.2

From this analysis we can conclude that all potential factors contribute to the estimate of the short interval. We can now do the same analysis on the long interval, and determine what its duration depends on. Table 3 shows the final model that came out of that analysis. The general pattern is the same as for the short interval: previous estimates of the long interval and previous feedback on that interval affect the current estimate, even longer back than for the short interval. This is probably due to the fact that the long interval changes. But also the estimate of the previous short interval and the feedback on that interval impact the next long estimate.

Model

The two models of which the results are shown in Figure 1 are in fact instantiations of the same model with different parameter settings. The basis for the model is two modules from the ACT-R theory (Anderson, 2007), but implemented in statistical package R (http://www.r-project.org/). More specifically, we used the time estimation modules (Taatgen, et al., 2007), and the declarative memory module augmented with the blending mechanism (Lebiere, Gonzalez, & Martin, 2007).

Time Estimation

The temporal module of ACT-R measures time in units that start at 100ms, but become gradually longer, creating a nonlinear representation of time. For the purposes of the

Table 3. Fixed effects in the regression model for the long interval

Fixed Effect	Value of β	*t* value
Intercept	695 ms	3.8
$long_{n-1}$	0.34	8.5
$long_{n-2}$	0.16	4.0
$long_{n-3}$	0.12	4.6
$long_{n-4}$	0.05	1.9
long-fb-S_{n-1}	159 ms	4.8
long-fb-L_{n-1}	-118 ms	-2.5
long-fb-S_{n-2}	82.9 ms	2.5
long-fb-L_{n-2}	3.8 ms	0.1
$short_{n-1}$	0.15	2.9
short-fb-S_{n-1}	85 ms	2.1
short-fb-L_{n-1}	-107 ms	-6.5

present model, the nonlinearity is not very important. The temporal module can be given a start signal, which resets the clock, after which an accumulator starts collecting pulses. The short interval of 2 seconds corresponds to approximately 17 pulses, and the long interval of 3.1 seconds to approximately 26 pulses. Noise is added to each pulse, which means that estimates are always approximate. For the purposes of the model, the important aspect of the time estimation module is that it can estimate a particular time interval by translating it into number of pulses, and that it can reproduce a time interval by waiting until a particular number of pulses has been accumulated. The noise produces variability in the estimates that correspond to variability in human time estimation.

Declarative Memory

The assumption of the model is that when a particular time interval has to be produced, the number of pulses representing that interval is retrieved from memory. There is no single representation of a particular interval in memory, but rather a collection of past experiences. Each past experience is represented by a memory chunk, which contains the type of interval (long or short), and a number of pulses. When an interval is retrieved from memory, each chunk receives an activation value on the basis of its age (how old is the experience), and whether it matches the current request:

$$A(t) = \log(t - t_{creation})^{-d} + mismatchpenalty$$

In this equation, $t_{creation}$ is the time the chunk is created, so the activation of a chunk decreases with time. The mismatchpenalty of a chunk is 0 if the request matches the chunk (e.g., we are retrieving a short interval and the chunk represents the short interval), but a negative value in the case of a mismatch (e.g., we try to retrieve a short interval but the chunk represents a long interval).

In standard ACT-R, activation determines the probability of retrieval of a chunk. This means that more recent experiences that match the request have the highest probability to be retrieved. The following equation estimates these probabilities (where *t* is a noise parameter, and the summation is over all candidate chunks):

$$P_i = \frac{e^{A_i/t}}{\sum_j e^{A_j/t}}$$

With the blending mechanism (Lebiere et al., 2007), however, a weighed average of all candidate chunks is retrieved. If we try to retrieve the duration of the short interval, the results will be a blend of all intervals in memory, with the more recent intervals having a higher impact, and the intervals that match the request (short) having a higher impact than the mismatching long intervals. The resulting value can simply be calculated by multiplying the number of pulses in a chunk (V_i) by the probability of retrieval:

$$\text{Result value} = \sum_j P_j V_j$$

In order to determine how many pulses to wait for an interval, the model not only retrieves the representation of the interval, but also feedback received for that interval. For this we use exactly the same mechanism as for the retrieval of the interval. Whenever feedback is received, the model stores this in memory. If the feedback was "correct" it stores the value of 0, if it was "too long" it stores a negative value, and when it is "too short" it stores a positive value (this value is referred to as the feedbackshift, which is a free parameter in the model). Retrieval is done in the same way as the retrieval of the interval itself. This means that the feedback of previous trial for the same duration has the highest impact, but that earlier feedback and feedback for the other duration can also weigh in.

To summarize: if the model has to produce a certain interval, it determines the number of pulses by retrieving a blend of memory representations for that interval. It then retrieves previous feedback for that interval, which is also a blend of earlier feedback. It adds the two together, and waits for that many pulses to produce the interval.

Table 4. Free parameters in model A and B

Parameter	Model A	Model B
Noise parameter *t*	0.25	0.2
Mismatch penalty between short and long for interval retrieval	-1.3	both
Mismatch penalty between short and long for feedback retrieval	-0.8	-0.92
Feedbackshift: how many pulses to add or subtract on the basis of feedback	8	1.8

The free parameters for model A and B were set to the values in Table 4. All other parameters were set to their ACT-R or time estimation module defaults (d=0.5, t_0=100 ms, a=1.02, b = 0.015). The parameters in model A were determined using the procedure that many modelers follow: starting with some initial set of parameters try varying them in order to optimize the fit in terms of R^2 and RMSE. This is typically a satisficing procedure (unless the whole parameter space is explored): model fitting ends as soon as variation of parameters leads to little improvement, and the current fit is decent enough. For model B we used a different method that we will outline later.

So Which is the Better Model?

When we create a cognitive model, it is not our goal to fit a particular data graph, although this may be part of the process, but to explain the phenomena that we are interested in. The statistical analysis has revealed that both the representations of the two intervals and the feedback for the intervals play a role in producing the next interval. It does not tell us what cognitive mechanisms can produce this. The cognitive model does supply a possible answer: a single memory mechanism that has been validated in many other studies can incorporate all factors that play a role in producing the estimate. But is this really true? The graphs in Figure 1 show a good fit, and the R^2's and RMSE also look decent, so what else is there to say?

We can test the impact of the factors that turned up significantly in the data more directly by performing the same analysis on the model outcomes. Statistical significance is not very relevant here, because we can run the model as often as we like. But the model should produce β values that are comparable to the β's found in the data analysis. We therefore ran each model 100 times, and collected the model data in the same format as the human data. This allowed us to fit the same linear regression models. Table 5 shows the results for two models next to the data.

On the basis of this analysis a whole new picture emerges: Model A does not fit the data at all, while Model B provides a very decent fit. The table also reveals the problem of Model A: its representation of the interval is much too stable, as is shown by the estimates for the intercept. In Model A, the intercepts are approximately equal to the actual duration of the interval, and there is hardly any impact of previously produced intervals, either long or short (as evidenced by the low $long_{n-x}$ and $short_{n-x}$ effects). Moreover, Model A's responses to feedback are much stronger than in the data. For example, if Model A receives the "too short" feedback on the short interval, it will respond to this by increasing its next production of that interval by 487 ms, while subjects only increase it by 110 ms. It probably needs such strong values to produce the shifts in estimates of the long interval.

Table 5. Comparison between model and data for the two models

Short interval			
Fixed Effect	β data	β Model A	β Model B
Intercept	657 ms	2157 ms	789 ms
$short_{n-1}$	0.385	0.08	0.356
$short_{n-2}$	0.085	-0.03	0.048
short-fb-S_{n-1}	110 ms	487 ms	170 ms
short-fb-L_{n-1}	-208 ms	-521 ms	-153 ms
$long_{n-1}$	0.16	-0.06	0.15
long-fb-S_{n-1}	92.6 ms	432 ms	125 ms
long-fb-L_{n-1}	-163 ms	-534 ms	-211 ms
Long interval			
Fixed Effect	β data	β model A	β model B
Intercept	695 ms	3162 ms	493 ms
$long_{n-1}$	0.34	0.011	0.22
$long_{n-2}$	0.16	0.012	0.25
$long_{n-3}$	0.12	0.003	0.12
$long_{n-4}$	0.05	0.001	0.09
long-fb-S_{n-1}	159 ms	626 ms	198 ms
long-fb-L_{n-1}	-118 ms	-744 ms	-251 ms
long-fb-S_{n-2}	82.9 ms	60 ms	90 ms
long-fb-L_{n-2}	3.8 ms	-142 ms	-57 ms
$short_{n-1}$	0.15	-0.07	0.18
short-fb-S_{n-1}	85 ms	326 ms	20 ms
short-fb-L_{n-1}	-107 ms	-492 ms	-35 ms

To summarize, Model A might produce a good global model fit, but for the wrong reasons. Model B on the other hand has factor values that are quite similar to those in the data. This means that the same factors that play a significant role in subjects' performance also play approximately the same role in the model's performance. This also means that it is reasonably likely that the model will generalize to other situations in which time intervals have to be stored in memory (see Note at the end).

In fact, the parameter settings for model B were derived by using the factors in the statistical model as an optimization criterion instead of the R^2 and RMSE values. Starting with model A, it was clear the feedbackshift had to be adjusted to reduce the factors associated with feedback. After that, some smaller adjustments led to model B.

Conclusions

Although there are several proposals to improve the assessment of model fit (e.g., Pitt et al., 2006; Weaver, 2008), not all of them are applicable to all types of models, and some of them require intensive additional calculations. The method we showed here is relatively straightforward in comparison, because the same method that is used to analyze the data (which has to be done anyway) can also be used to analyze the model's fit. Although this comparison does not produce a nice and simple single value for the quality of the fit, such a value might be an illusionary

concept anyway. It is never possible to prove that a model has "a 95% probability of being correct". For this it is necessary to know the complete space of possible models/theories, something that is decidedly undecidable.

The nice thing about this analysis is that we can see whether the model produces the effects that we are interested in, and that it produces them in approximately the same order of magnitude. It was even helpful in data fitting itself, because it shows what particular factor is throwing the fit out of balance.

In conclusion, analyzing model fits with mixed-effect models is a promising tool in the modeler's toolbox.

Note

The experiment that we have discussed here had two additional conditions, one in which both intervals remained constant for the duration of the experiment, and one in which they long interval became shorter first and longer later. The model we presented here has not run for those conditions yet. We will do so before the conference and present the results there, and we will keep our fingers crossed that the fit will be good.

Acknowledgements

We would like to thank the participants of the Cognitive Modeling class in Groningen in participating in the discussion of these data, and Stefan Wierda for collecting the data.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford university press.

Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, *35*, 54.

Lebiere, C., Gonzalez, C., & Martin, M. (2007). Instance-based decision making model of repeated binary choice. In *proceedings of the 8th International Conference on Cognitive Modeling*. Ann Arbor, Michigan, USA.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113*, 57–83.

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

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded Cognition: An Integrated Theory of Concurrent Multitasking. *Psychological Review, 115*(1), 101-130.

Schunn, C. & Wallach, D. (2005). Evaluating goodness-of-fit in comparison of models to data. In: W. Tack (Ed.), *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack* (pp. 115-154). University of Saarland Press, Saarbrücken, Germany.

Taatgen, N. A., Huss, D., Dickison, D. & Anderson, J. R. (2008). The acquisition of robust and flexible cognitive skills. *Journal of Experimental Psychology: General*, *137*(3), 548-565.

Taatgen, N. A., van Rijn, H., & Anderson, J. (2007). An Integrated Theory of Prospective Time Interval Estimation: The Role of Cognition, Attention, and Learning. *Psychological Review, 114*(3), 577-598.

van Rijn, H. & Taatgen, N.A. (2008). Timing of multiple overlapping time intervals: How many clocks do we have? *Acta Psychologica, 129*(3), 365-375.

Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: a statistical view informed by ACT-R. *Cognitive Science, 32,* 1349-1375.