Language &
Cognitive
Dynamics
Laboratory

Technical Report 2012.01
Version 1.0: June 2012

What is "Dynamic Consistency" and does it matter?

Dan Mirman

Abstract

Regression methods for analysis of fixation time course data differ in terms of their functional form, with polynomial regression and non-linear regression being the two main categories. Non-linear regression models are better-suited to capturing the asymptotic portions of fixation proportion curves, but are not *dynamically consistent*: the model of the average is not equal to the average of the individual models. In this report, lack of dynamic consistency is demonstrated concretely with analysis of data from a simple semantic competition "visual world paradigm" experiment. Visual and quantitative analysis reveals why lack of dynamic consistency undermines statistical inference based on central tendencies. The analyses also bring into question previous claims that non-linear models yield "intuitively" interpretable parameters. Dynamically consistent functional forms (such as polynomials) used in the context of multi-level regression appear to be the most promising analysis tools for fixation time course data.

What is "Dynamic Consistency" and does it matter?

More and more researchers are adopting regression methods for analysis of eye tracking (and other time series) data. These methods can be grouped into two broad categories: polynomial regression (Barr, 2008; Mirman et al., 2008) and non-linear regression[1] (e.g., McMurray et al., 2010; Scheepers et al., 2008). Non-linear regression models are better-suited to capturing the asymptotic portions of fixation proportion curves, but are not *dynamically consistent*: the model of the average is not equal to the average of the individual models. At the heart of much of statistical inference in experimental psychology is the assumption that the mean represents the central tendency of a sample of individual observations. As a result, violating this assumption could undermine the very foundation of how we think about our data and make inferences about it. To illustrate what this means, a prominent non-linear regression method was applied to data from a simple semantic competition experiment (Mirman & Magnuson, 2009) that tested a relatively large sample (*N* = 38) from a relatively homogenous population (undergraduate college students at the University of Connecticut).

Methods

Data were drawn from a simple VWP experiment in which listeners were more likely to fixate an object that was semantically related to the target than an unrelated object (for details of experiment design see Mirman & Magnuson, 2009). On each trial, the 4-picture display included a target object and a critical distractor, which was either semantically related or unrelated to the target. For simplicity, only competitor fixations for the "near" (strongly related) competitors and the matched unrelated distractors are considered here. Relatedness pairings were counterbalanced across participants, so each of the 38 participants saw each target only once with either a related or unrelated distractor, but each target appeared with both distractors across participants. The time course was divided into 13 100ms time bins starting at word onset. Terminated trials were considered as non-object fixations.

Mirman and Magnuson (2009) analyzed these data using polynomial regression with fourth-order orthogonal polynomials and found statistically reliable effects of semantic relatedness on all model terms, particularly on the intercept and quadratic terms. Here those data are re-analyzed with a non-linear regression approach using a logistic power peak (LPP) function (Scheepers et al., 2008):

$$fix\left(t\right) = \alpha\left(1 + \exp\left(\frac{t + \beta\ln\left(\gamma\right) - \delta}{\beta}\right)\right)^{\frac{-(\gamma+1)}{\gamma}} * \exp\left(\frac{t + \beta\ln\left(\gamma\right) - \delta}{\beta}\right) * \left(\gamma + 1\right)^{\frac{(\gamma+1)}{\gamma}}$$

where *α* is the amplitude, β is the width, *δ* is the location of the peak, and *γ* is the symmetry of the fixation proportion curve. First, the mean fixation proportion curve for each condition was computed and fit using the LPP function. This yielded 4 parameters for each of the 2 conditions. A Wald test was used to independently compare each parameter pair; that is, to assess whether each parameter estimate (*α*, β, *δ*, or *γ*) was statistically significantly different between the Related and Unrelated conditions. Then this model-fitting procedure was repeated separately for each participant, producing 304 total parameters (38 participants * 2 conditions * 4 parameters), which could be compared with

[1] We are specifically referring to non-linear models that are not dynamically consistent, which includes the non-polynomial models that have been used in analyses of VWP data.

paired-samples t-tests. The LPP model was fit by least-squares estimation using the *nls* function in R version 2.13 (R Development Core Team, 2011).

Model fits

The LPP model provided good fits separately for the averaged data and to each individual participant's data (see Supplemental Figure for individual participant model fits). For 6 of the 76 individual curves, the model fitting procedure did not fully converge (Related condition: participants #2, 10, 12, 31, and 36; Unrelated condition: participant #8); though incomplete convergence was not associated with poorer model fit (i.e., the fully converged models did not have lower overall deviance; $p > 0.25$). Figure 1 shows the average observed data (symbols), the model fit to the average data (solid black lines), the model fits for each individual participant (grey lines), and the model created by averaging the parameters of the individual participant models (all individuals: dashed lines; only fully converged models: dotted lines).

This graphical representation illustrates the consequences of lack of dynamic consistency: (1) the average of the individual models does not match the model of the average (solid and dashed black lines do not match up), and (2) the average of the individual models does not reflect the central tendency of the individual models themselves (i.e., the dashed and dotted black lines do not follow the thin gray lines).

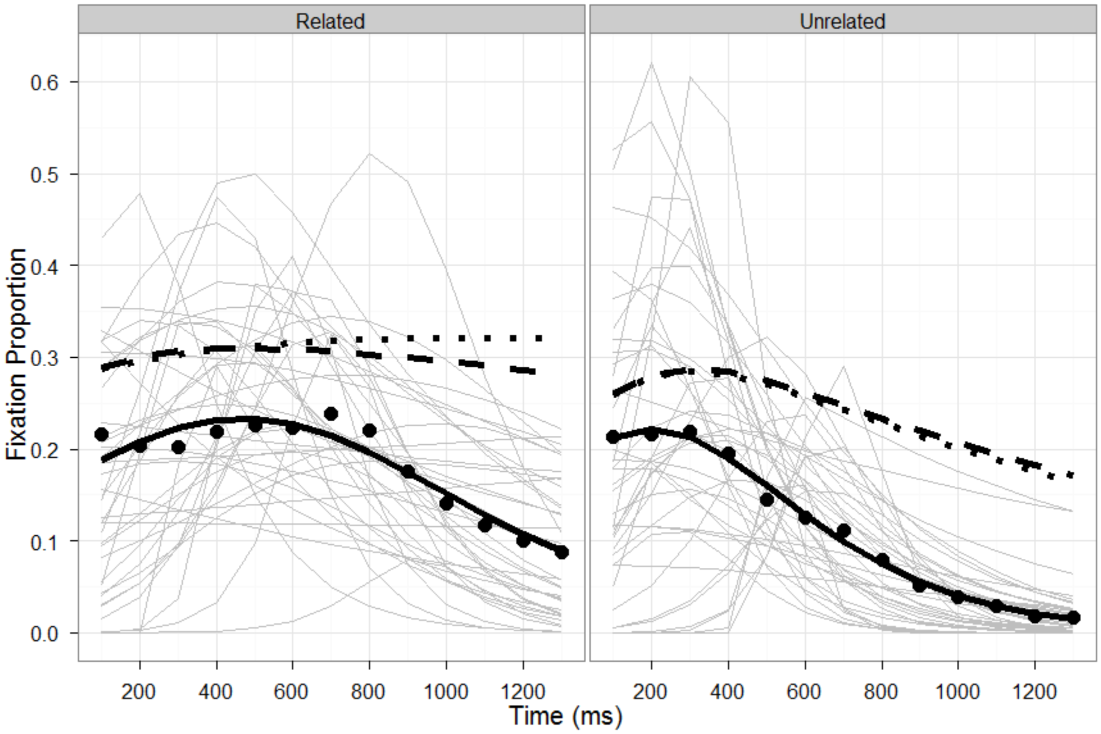


Figure 1. Comparison of model fits. The two panels correspond to the two distractor conditions (Left: Related; Right: Unrelated). The symbols indicate the mean of the gaze data. The thick black lines correspond to the three model fits: the model of the average data (solid lines), the average of all of the individual models (dashed lines), and the average of just the fully converged models (dotted lines). The thin gray lines show the individual participant model curves.

In other words, dynamically inconsistent models such as the LPP function are *non-linear in their parameters*, so the (linear) mean of those parameters does not reflect their central tendency. In contrast, polynomial functions are curvilinear in shape, but linear in their parameters; thus, a multi-level modeling approach can capture both the overall average data and the individual participant data[2]. This qualitative difference is discussed quantitatively in the next section.

Statistical comparisons

Table 1 shows the parameter estimates from the model of the average data, the average of all individual participant models, and the average of just the fully converged models. For the model of the average, a Wald test revealed a difference between conditions only for the location parameter δ ($\chi^2(1) = 4.96$, $p < 0.05$; all other $p > 0.25$). Paired-samples t-tests on the parameters from only the converged models revealed significant condition differences for the location parameter δ ($t(31) = 3.1$, $p = 0.0039$), a non-significant trend for the width parameter β ($t(31)=1.7$, $p = 0.10$), and no other differences (all other $t < 1$). When all participants were included in this analysis, the location parameter effect was no longer statistically reliable and the width parameter was marginally different ($t(37) = 1.8$, $p = 0.076$); the reliability of other effects was unchanged (see Table 1 for statistical results from all comparisons).

Table 1. Parameter estimates (SE in parenthesis) for LPP models.

	Model of the average			Average of all individual models (*N* = 38)			Average of fully converged individual models (*N* = 32)		
	Related	Unrelated	Wald test	Related	Unrelated	t-test	Related	Unrelated	t-test
α (amplitude)	0.234 (0.01)	0.221 (0.005)	$\chi^2(1) = 1.16$, *n.s.*	0.321 (0.20)	0.284 (0.02)	$t = 1.2$, *n.s.*	0.310 (0.18)	0.288 (0.024)	$t < 1.0$, *n.s.*
β (width)	3.90 (1.73)	2.34 (0.62)	$\chi^2(1) = 0.85$, *n.s.*	5.41 (2.01)	1.69 (0.21)	$t = 1.8$, $p < 0.1$	2.45 (0.39)	1.64 (0.214)	$t = 1.7$, $p = 0.1$
γ (symmetry)	1.00 (1.13)	1.27 (0.54)	$\chi^2(1) = 0.22$, *n.s.*	23.0 (12.3)	8.89 (4.72)	$t = 1.0$, *n.s.*	25.8 (14.5)	9.66 (5.57)	$t < 1.0$, *n.s.*
δ (location)	4.68 (0.47)	2.01 (0.27)	$\chi^2(1) = 4.96$, $p < 0.05$	10.4 (4.96)	3.00 (0.29)	$t = 1.5$, *n.s.*	4.69 (0.46)	3.07 (0.319)	$t = 3.1$, $p < 0.01$

As depicted graphically in Figure 1, the parameter estimates for the model of the average and the average(s) of the individual models in Table 1 are not the same, representing a concrete example of lack of dynamic consistency. It should now be clear why lack of dynamic consistency undermines the interpretation of the paired-samples t-test results: the t-test told us that the *mean* of the individual participants' location parameter δ's were unlikely to be the same for the two conditions, but those means are not the same as the parameters for the model of the average.

Note also that the location parameter is the least intuitive of the four parameters for capturing the effect of semantic competition; amplitude or width would map more intuitively to theories of semantic activation, which predict greater and longer-lasting activation of semantically related concepts than unrelated concepts during spoken word comprehension. This is relevant because advocates of non-

[2] Multi-level regression can be used with any function that is linear in its parameters, i.e., dynamically consistent, not just polynomial functions. Any "non-linear" functional form that can be implemented in a multi-level regression model would, by definition, be dynamically consistent and exempt from the concerns raised here.

linear models sometimes claim that the parameters have intuitive mappings to psychological processes. On this view, these results would argue that semantically related items are activated later than unrelated items, but not activated more or for a longer period of time, which conflicts both with theories of word comprehension and visual intuitions from looking at the data. The point is that unless the experimenters have set out to test an explicit mathematical model of the underlying cognitive processes, statistical models should be chosen for their ability to quantify the observed data and not for purportedly "intuitive" mappings to psychological processes.

Convergence issues

In addition to lacking dynamic consistency, non-linear models also seem not to converge reliably for individual participant data. In the above example the problem was not too pervasive – less than 8% of the models (6 out of 76) failed to fully converge. However, researchers who have used non-linear regression methods to analyze gaze data have specifically pointed out that analysis by participants or items was not feasible (e.g., Scheepers et al., 2008, p. 18 footnote 9) and opted for analyzing overall average data or sub-sample data. Some (Apfelbaum et al., 2011; McMurray et al., 2008) have approached this problem using *jackknifing*, which reverses the standard process of evaluating between-participant variance. Instead of selecting each individual participant and fitting a model to his or her data, each participant is sequentially excluded from analysis of the remaining participants' average data. For an experiment with N participants, this produces N sets of parameters, each based on average data from N-1 participants. The variance estimate is then adjusted for the fact that each participant was included in N-1 sub-samples.

When data are sparse or noisy, statistical resampling techniques such as jackknifing are a common approach, and this has been articulated as the motivation for using them with gaze data, even when each participant completes a relatively large number of trials per condition (e.g., 36 trials per condition in Apfelbaum et al., 2011). A LPP regression jackknife analysis of the example data used here converged for all subsamples and, not surprisingly, the dynamic inconsistency problem was reduced: the averages of the individual parameter estimates were very similar to the parameter estimates of the model of the average data (less than 1.5% difference for each of the 8 parameters); though the fact that the differences were not eliminated indicates that dynamic inconsistency is intrinsic to the functional form. Statistical analysis of these parameter estimates was consistent with the other LPP-based analyses: using the appropriately adjusted error term, the only reliable difference was on the location parameter ($t(37) = 2.96$, $p < 0.01$; all other $t < 1$, $p > 0.15$).

Individual differences

An additional challenge for non-linear regression approaches is that dynamic inconsistency precludes analysis of individual differences. Researchers are increasingly interested in using eye tracking to analyze individual differences (e.g., McMurray et al., 2010; Kalenine et al., 2012; Mirman et al., 2008; Mirman et al., 2011), which is a natural extension of multi-level polynomial regression, but appears problematic for non-linear models.

Conclusions

In sum, lack of dynamic consistency poses a serious problem for non-linear regression models of gaze data. The logic of comparing sample means is invalid when the model of the average is not equal to the average of the individual models. Analyzing only overall average data can avoid this problem, but this approach conflates the sample mean and the population mean (i.e., the variability across individuals is not considered when evaluating condition differences). Analyzing subsamples, as in jackknifing or other resampling techniques, may reduce the severity of dynamic inconsistency while including individual participant variability in the analysis, but this approach precludes analysis of individual differences. Multi-level modeling resolves these problems by simultaneously capturing both the overall average data and the individual-level deviations from the average. Polynomial regression can be readily implemented in a multi-level modeling framework for analysis of gaze data (e.g., Barr, 2008; Mirman et al., 2008), but any dynamically consistent functional form can be used in a multi-level modeling framework.

References

Apfelbaum, K. S., Blumstein, S. E., & McMurray, B. (2011). Semantic priming is affected by real-time phonological competition: Evidence for continuous cascading systems. *Psychonomic Bulletin & Review, 18*(1), 141-149.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*(4), 457-474.

Kalénine, S., Mirman, D., & Buxbaum, L. J. (2012). A combination of thematic and similarity-based semantic processes confers resistance to deficit following left hemisphere stroke. *Frontiers in Human Neuroscience*, *6*(106).

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review, 15*(6), 1064-1071.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology, 60*(1), 1-39.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59*(4), 475-494.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition, 37*(7), 1026-1039.

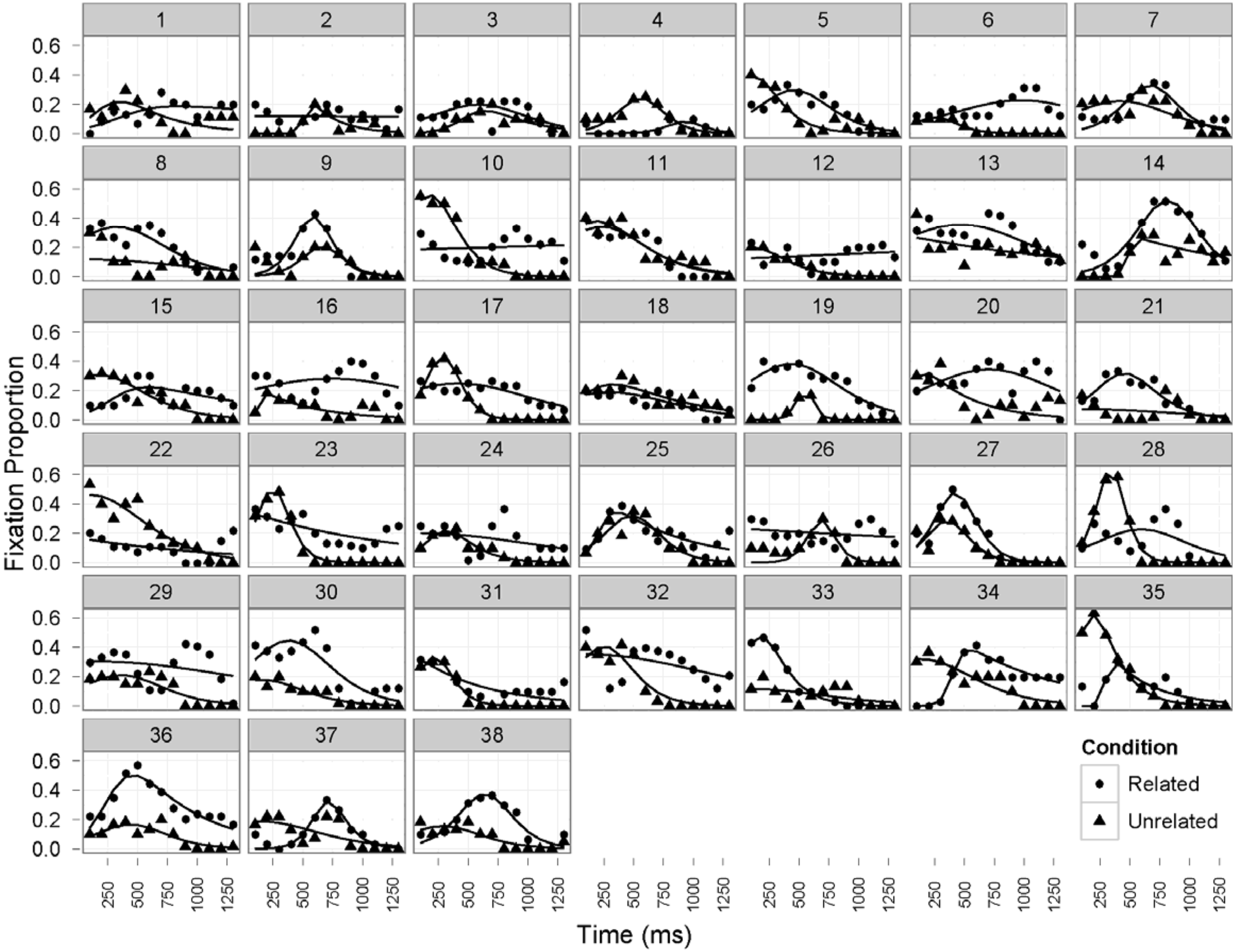
Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language, 117*(2), 53-68.

R Development Core Team (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Scheepers, C., Keller, F., & Lapata, M. (2008). Evidence for serial coercion: A time course analysis using the visual-world paradigm. *Cognitive Psychology, 56*(1), 1-29.

Skovlund, E., & Fenstad, G. U. (2001). Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? *Journal of Clinical Epidemiology, 54*, 86-92.

Supplemental Figure: Individual data and LPP model fits


Fixation Proportion
Time (ms)
Condition
● Related
▲ Unrelated