Modelling mechanisms of persisting and resolving delay in language development

Michael S. C. Thomas & V. C. P. Knowland

Supplementary material

1. Simulation details

1.1. Base model

A 3-layer, backpropagation network was used to learn to output the past-tense form of a verb from an input vector combining a phonological representation of the verb stem and lexical-semantic information (Joanisse & Seidenberg, 1999). The architecture is shown in main article, Figure 1. The training set was the "phone" vocabulary from Plunkett and Marchman (1991, p. 70). This comprised an artificial language set constructed to reflect many of the important structural features of English past-tense formation. There were 500 monosyllabic verbs, constructed using consonant-vowel templates and the phoneme set of English. Phonemes were represented over 19 binary articulatory features, a distributed encoding based on standard linguistic categorisations (Fromkin & Rodman, 1988). Separate banks of units were used to represent the initial, middle, and final phonemes of each monosyllable. The output layer incorporated an additional 5 features to represent the affix for regular verbs. The input layer included 500 units to encode the lexical status of each verb in the training set using a localist encoding scheme (Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003). Networks thus had 3x19+500=557 input units and 3x19+5=62 output units. There were four types of verbs in the training set: (1) regular verbs that formed their past tense by adding one of the three allomorphs of the +ed rule, conditioned by the final phoneme of the verb stem (e.g., *tame-tamed*, *wrap-*

wrapped, *chat-chatted*); (2) irregular verbs whose past-tense form was identical to the verb stem (e.g., *hit-hit*); (3) irregular verbs that formed their past tenses by changing an internal vowel (e.g., *hide-hid*); (4) irregular verbs whose past-tense form bore no relation to its verb stem (e.g., *go-went*). There were 410 regular verbs, and 20, 68, and 2, respectively, of each irregular verb type. A separate set of novel verbs was constructed to evaluate the generalisation performance of the network. These verbs could differ depending on their similarity to items in the training set. Generalisation in this case was assessed via 410 novel verbs each of which shared two phonemes with one of the regular verbs in the training set, and was evaluated based on the proportion of these novel verbs that were assigned the correct allomorph of the regular past-tense rule.

1.2. Encoding extrinsic variation

Each network simulated a child raised in a given family, and families were assumed to vary in the richness of the language used. The language input was assumed to vary to some extent according to SES (Hart & Risley, 1995). A training set was created for the past-tense information available in each family environment. SES was implemented through generating a *family quotient* for each simulated child. The family quotient was a number between 0 and 100%. This value was used as a probability determining whether each verb in the perfect training set would be included in the family's vocabulary. The family training set was then fixed throughout development. Performance was always assessed against the full perfect training set (analogous to a standardised test of past-tense formation applied to all children). The family quotient manipulation corresponded to a reduction in type frequency for both regular and irregular verbs. Based on the findings of Thomas, Ronald and Forrester on the appropriate range of intrinsic versus extrinsic variation to capture data on past tense acquisition, family quotients were sampled from a uniform distribution from 60% to 100%, corresponding to learning environments with reasonably high quality. This corresponds to the assumption that there is a minimum amount of linguistic information typically available to a child.

1.3. Encoding intrinsic variation

Connectionist networks contain a range of parameters that increase or decrease their ability to learn a given training set. Parameters such as learning rate, momentum, and number of hidden units feature in most published simulations. In models of normal/average development, parameters are optimised to achieve best learning (usually in the presence of the perfect training set). In the current model, a number of parameters were simultaneously varied across individual networks, with learning ability determined by their cumulative affect. The mechanistic variations producing differences in the rates of development were therefore only *quantitative*. Variations occurred over fourteen computational parameters, allowing for over 2000 billion unique individuals.

The parameters were as follows: *Network construction*: Architecture, number of hidden units, range for initial connection weight randomisation, and sparseness of initial connectivity between layers. *Network activation*: unit threshold function, processing noise, and response accuracy threshold. *Network adaptation*: backpropagation error metric used in the learning algorithm, learning rate, and momentum. As well as an overall learning rate, there were separate parameters modifying the learning rate between the semantic input units and the hidden units, and the phonological input units and the hidden units, potentially altering the relative balance of these sources of information during learning, and therefore allowing more lexical or phonological strategies to past-tense acquisition. *Network maintenance*: weight decay, pruning onset, pruning probability, and pruning threshold.[1]

These parameters can be viewed as serving different types of processing role within the network, although some parameters contribute to more than one role. Some parameters alter the network's learning *capacity*, that is, the complexity and the amount of information that can be learned. These include the architecture, the number of hidden units, and the initial sparseness of connectivity. *Regressive events* involving pruning of connections can also reduce capacity later in development, implicating the pruning onset, pruning probability and pruning threshold parameters in predicting learning trajectories (see Thomas, Knowland & Karmiloff-Smith, 2011). The nature of the learning algorithm determines both what can be learned and also how quickly. The speed of learning can be thought of as the network's *plasticity*. Other parameters alter plasticity, including the learning rate parameter, the learning rates in semantic and phonological connections, the momentum, the initial range of weight variation, and the unit threshold function. The unit threshold function determines how responsive a processing unit is to variations in its input, and therefore to some extent determines the quality of the *signal* propagating through the network. Signal is also affected by the level of processing noise, and the accuracy required of output units to drive a response. Combined with the quality of the learning environment, the mechanisms affecting development can be broadly assigned the following four categories: *capacity*, *plasticity*, *signal*, and *environment*. Parameters are categorised in this way in the reporting of results.

[1] Formal specification of the parameters and their value ranges can be found in a technical report available at http://www.psyc.bbk.ac.uk/research/DNL/techreport/Thomas_paramtables_TR2011-2.pdf

1.4. Design

Development was traced across a population of 1000 simulated individuals, focusing on the rate of acquisition of regular English past-tense forms. One thousand sets of the 14 computational parameter values were generated at random, with parameters sampled independently (see footnote 1). These sets were instantiated as 1000 connectionist networks. A family quotient value was generated for each network and used to create an individualised family training set. Each network was trained for 1000 epochs on its family training set. At each epoch, performance was measured on the perfect training set. Performance was assessed on regular verbs, irregular verbs, and on generalisation of the past-tense rule to novel forms, in order to generate a behavioural 'profile' for each network. Performance was measured via accuracy levels (% correct).

Early performance on regular verb acquisition was used to define a delay group (see below) and their subsequent progress was then traced with reference to the population normal range. The robustness of the results was tested by three subsequent analyses: evaluating delay defined according to the acquisition of irregular (vowel-change) verbs; evaluating the outcome of early-identified 'gifted' performance, that is, outliers falling above the normal range rather than below it; and evaluating regular acquisition in a population of simpler networks which considered past tense as a mapping only between phonological forms and did not include lexical-semantic input. Results from these subsequent analyses are described to illustrate the robustness of the model but are not reported in detail.

2. Results

2.1. Comparisons of Time 1 profiles of persisting and resolving delay groups – further analyses

Time 1 performance contained some individuals with scores at floor; this may have differentially affected irregular verbs, which are harder to learn, producing the lack of a reliable difference between persisting and resolving delay profiles. The group comparison was carried out at each subsequent time point to evaluate this possibility. Of course, differences must increase on all measures, because by definition, the performance of persisting and resolving groups will diverge. However, if the initial overlap in irregular verb performance between the groups was due to floor effects, the effect size of the irregular difference should approach those of regular and novel verbs as performance comes off floor in those individuals. Figure s.1 plots the effect sizes of the profile differences between the groups at each time point. It demonstrates that the effect sizes were consistently larger for regular and novel verbs than for irregulars. Persisting and resolving groups, then, were distinguished at time 1 by small differences in the ability to abstract the regularities of past tense morphology from exposure to the learning environment, with the persisting group less able to do so.

====================
Insert Figure s.1 about here
====================

One might ask whether this result was an inevitable consequence of using past-tense acquisition as our model domain, and of defining delay based on the acquisition of regular verbs. Figure s.1 also contains an equivalent comparison between two groups of individuals who were defined as 'gifted' at time 1 (i.e., their performance fell more than 1 standard deviation *above* the population mean). Of these early gifted individuals, 70.0% (142) returned to the normal range by time 5, leaving

30.0% (61) individuals who exhibited a sustained pattern of giftedness. The time 1 profiles for delay and giftedness demonstrated a contrasting pattern: sustained and non-sustained giftedness were distinguished by small initial differences in performance on regular and irregular verbs, but not on novel verbs. These differences index speed of acquisition of the training set, rather than extracting regularities that could be generalised to novel cases. The contrasting pattern between delay and giftedness suggests that the ability to extract regularities was a particular predictor of delay outcome in this system, not an inevitable consequence of analysing the tails of the distribution for networks exposed to the past-tense problem domain.

2.2. MANOVA and Regression analyses to identify neurocomputational parameters that distinguished delay groups

Table s.1 shows the results of a statistical comparison of the mean neurocomputational parameter values for simulated typically developing, persisting delay and resolving delay groups. It contains the results of two complementary analyses: multivariate analysis of variance and multinomial logistic regression (a third method, linear discriminant analysis was also available, but we do not report these results because the weighted linear combination of parameters that it produces is a misleading depiction of the non-linear way in which parameters interact in the model). Table s.2 shows equivalent results for a comparison of the resolving delay group, split by whether the final outcome was low (bottom 500 of population), good (top 500), or very good (top 200). Table s.3 incorporates the mean parameter values per group, as well as three case studies of individual parameter sets, which are discussed below.

====================

Insert Tables s.1-3 about here

====================

2.3. Case studies of simulated persisting and resolving delay

Table s.3 contains group means of the neurocomputational parameter sets of artificial neural networks in the typically developing, persisting delay, and resolving delay groups respectively. Such group means risk being opaque or even misleading, in that they may not represent the parameter set of any actual individual in the population. In this sense, case studies can serve a complementary role. Three case studies are also included in Table s.3, one exhibiting persisting delay (PD1), one exhibiting resolving delay with low final outcome (RD1), and one exhibiting resolving delay with very good final outcome (RD2). The parameter sets of these individual networks and how they generated the outcomes are discussed below. A consideration of each parameter set must first explain first why the individual should show slow early development and then in two cases, why this delay should resolve.

Case PD1 had limited computational capacity in its two-layer architecture, sparser connectivity, and a less powerful learning algorithm. Three parameters contributing to plasticity were also reduced. However, there was a rich learning environment. PD1 was a case of poor outcome through reduced capacity and plasticity in the face of rich information.

Case RD1 had fewer than normal hidden units, but a more powerful architecture and learning algorithm than PD1. It had two parameters contributing to low plasticity, and a signal property parameter requiring accurate outputs to drive responses. These accurate outputs take longer to acquire during training. In addition, processing noise was elevated. The environment was of average quality. Early delay occurred due to lower plasticity and signal problems, but adequate computational power and an average environment led to a low-normal outcome.

Case RD2 was similar to RD1, with adequate computational capacity but low plasticity and signal issues. The environment, however, was much richer, allowing for a high final performance level.

Notably in these three case studies, *no single parameter was the cause of the delayed developmental trajectory*, rather several parameters interacted. We see the interplay between four effects of capacity, signal quality, plasticity, and the environment. (Not included here is the possibility of early regressive events that reduce capacity.) Poor capacity and plasticity were associated with poor outcome independent of the environment. Low plasticity and signal problems could be overcome with extended training, with the final outcome dependent on the quality of the environment.

2.4. Details of the Bishop (2005) study, the data of which allowed a comparison of SES effects on the outcome of persisting versus resolving delay groups

Bishop (2005) analysed data from the large British sample of twins considered in Dale et al. (2003) and Bishop et al. (2003). Bishop (2005) identified a sample of the twins who exhibited language delay risk at 4 years of age. On the basis of parental report at 4 years of age (Dale et al., 2003), children were identified as at risk of language impairment if they had a poor score on any one of three indices: (1) child described as not yet talking in full sentences, (2) vocabulary was in the lowest 10% based on parental checklist, or (3) parents answered ‘‘yes’’ to the question ‘‘Do you have any concerns about your child’s speech and language?’’ and selected the option ‘‘his/her language is developing slowly’’ when indicating the nature of the concern. These children, along with a sample of twins not identified as at risk, were tested at 6 years of age on a test of English past tense production (the Past Tense probe subtest of the Rice-Wexler Test of Early Grammatical Impairment, 2001). At 6 years of age, around one third of the early language impairment risk group then met psychometric criteria for SLI, compared to one in ten of those not identified as at risk (Bishop et al., 2006). Once more, these data illustrate the familiar pattern, with a high proportion of resolving early delay. The final sample included 442 6-year-old children, for whom SES information was also available (Bishop, 2005; Petrill et al., 2004; see Thomas, Ronald & Forrester, submitted, for further details of the full sample). Here, we focus on those children identified by early markers of delay, and consider the relationship between the SES marker, and performance on English past-tense formation. Past-tense data were available for 94 6-year-old children both exhibiting language impairment risk at 4 years and meeting psychometric criteria for SLI at 6 years, for 104 6-year-old children exhibiting language impairment risk at 4 years of age but not meeting psychometric criteria for SLI at 6 years, and 166 children exhibiting neither language impairment at 4 nor SLI at 6.

The marker of SES was derived in the following way. Demographic information was obtained via questionnaire from the first contact with the family at age 4 and included five pieces of information. These were the father’s highest educational level and occupational status, the mother’s highest educational level and occupational status, and the age of mother at birth of the eldest child. From these data, an index of SES was created based on a factor analysis (Petrill et al., 2004, p.448). This method yielded a scale ranging from -1.57 (low SES) to +1.54 (high SES), with a mean of -0.16 and a standard deviation of 0.72. Figure 5, main article, shows a comparison of the mean SES scores for typically developing, persisting, and resolving groups, as well as simulation data (SES scores and FQ parameter values have been rescaled to a common scale for ease of comparison). As with the simulation data, both early delay groups combined yielded reliably lower SES scores than the typically developing group ($t(362)=2.75$, $p=.006$, Cohen's $d=.290$). In this case, the persisting group was also reliably lower in SES compared to the typical group, but no other differences were reliable (TD vs. PD: $t(258)=3.18$, $p=.002$; TD vs. RD: $t(268)=1.44$, p.152; PD vs. RD: $t(196)=1.60$, $p=.111$). As with the simulations, the SES measure demonstrated only a small ability to predict individual differences in regular past tense formation, explaining only 0.5% of the variance across all groups ($F(1,362)=1.85$, $p=.175$). In part, this was due to ceiling effects in regular verb scores.

3. Discussion

3.1. Evaluation of the model: Strengths

The model succeeded with respect to its four identified aims: (1) to establish whether a quantitative account of the developmental variations in a population is sufficient to generate subgroups demonstrating persisting delay and resolving delay, (2) identifying differences in behavioural profiles when delay is first diagnosed and evaluating their ability to predict developmental outcomes, (3) assessing the role of environmental variation in causing developmental delays or aiding their resolution, and (4) suggesting possible mechanisms responsible for producing cases of persisting versus resolving delay. It was successful with respect to several roles of computational modelling in developmental theory: demonstrating the viability of a theoretical proposal, unifying a range of empirical data, and generating novel testable predictions. Note, however, that the current model was not specifically designed to address the issue of delay, nor were its parameters tailored to capture the difference between persisting and resolving delay. The findings were emergent, in the sense that the current analysis took the simulated population of Thomas, Ronald and Forrester (submitted), which was addressed to the issue of how SES influences language development, identified simulated individuals exhibiting early delay and then traced developmental outcomes. The findings were a consequence of the way in which individual variation was encoded in this model of language development.

One might ask how robust the simulation findings were, given that only one population was simulated, and therefore only one proportion of persisting versus resolving delay cases was reported. Recall, the numbers of simulated individuals in each group were: typically developing (TD) = 713, persisting delay (PD) = 118, resolving delay low outcome (RDL) = 136, resolving delay good outcome (RDG) = 28, and resolving delay very good outcome (RDVG) = 5. Using the same population, early delay was instead defined on the basis of irregular vowel-change verb performance, and outcomes assessed. This gave the following proportions: TD=756, PD=170, RDL=63, RDG=10, RDVG=1. A different population was trained using a simpler past-tense architecture, which only employed phonological information (Plunkett & Marchman, 1991). With delay defined on regular verb performance, the proportions were as follows: TD=765, PD=64, RDL=127, RDG=37, RDVG=7. Finally, using the original population, individuals were assessed on the basis of scoring above the normal range early in development, and outcomes were once more traced. Although ‘giftedness’ (G) did not behave like delay in several respects (such as in the contributory role of intrinsic and extrinsic parameters), similar proportions of persisting and resolving patterns were found: TD=797, PG=61, RGH=134, RGL=6, RGVL=2. Overall, then, the qualitative pattern of results appeared to be robust over a number of conditions.

3.2. Evaluation of the model: Limitations

As with any implemented model, there were limitations. First, the model was used in a more illustrative setting as a notional developing system applied to a structured language domain, rather than as a specific model of the acquisition of inflectional morphology. This is because the target empirical phenomenon considered disparate measures of language ability for the early and later markers of language delay. In many cases delay was initially diagnosed with respect to vocabulary, or even global concerns of the parents regarding their children's language development, with later assessments considering a wider range of measures that sometimes included morphosyntax. In the illustrative setting, the model evaluated the hypothesis that patterns of persisting and resolving delay could emerge in the development of a single system. Nonetheless, a novel prediction generated by the model was subsequently tested against and supported by empirical data from inflectional morphology. Second, the past-tense model itself involved a number of simplifications necessitated by population modelling. For example, the model employed an artificial past tense like problem domain rather than a full-scale English verb corpus.

One of the simplifications of the model was that changes in the intrinsic properties of the artificial neural networks were restricted to learning properties, rather than the structure of the input and output representations. Some models have considered how differences in phonological or semantic information supplied to learning systems can cause variations in developmental trajectories, perhaps even simulating SLI (Hoeffner & McClelland, 1993; Karaminis, 2011; Thomas & Karmiloff-Smith, 2003). If the key distinctions required to learn the latent structure of a language domain are not present in the inputs and outputs presenting to a learning system, clearly acquisition cannot be successful. On the other hand, if those distinctions are simply encoded less saliently (that is, with smaller activation differences), learning may be ultimately successful but take longer. In other words, it is plausible that a distinction between persisting and resolving delay would occur if the manipulation of input and output representations were added to the parametric variations. Of course, this remains to be demonstrated with further simulations.

A second simplification of the model was that intrinsic and extrinsic parameters were sampled independently. This means that the model did not consider the possibility of gene x environment correlations, i.e., the possibility that there might be a correlation between individuals possessing suboptimal computational parameters in language-learning systems and their exposure to a poorer language environment. While no definitive empirical evidence has demonstrated a correlation between children possessing, say, gene variants that place them at risk of language disorders and being raised in a low SES family, such a correlation is implied now or in the future by the combined findings that (a) language disorders are heritable and (b) as adults, individuals with speech and language disorders have lower SES scores (e.g., Bishop, North & Donlan, 1995; Felsenfeld, Broen & McGue, 1994). Were the model to include a correlation between intrinsic and extrinsic parameters across individuals, this would have a greater impact on resolving delay (where a richer environment can produce good final outcomes) and a lesser impact on persisting delay (where the principal limitation stems from an intrinsic reduction in computational capacity).

Despite the empirical effects that were successfully simulated by the model, it could, of course, turn out to be wrong: implemented models serve only to demonstrate the viability of theoretical proposals, they cannot demonstrate that the truth of those proposals. Thus, it could be that persisting and resolving delay are qualitatively different, or that some instances of persisting delay are qualitatively different. It could also be that a proportion of the cases of resolving delay arise from the way early delay is diagnosed, so that early assessments of vocabulary bring together a heterogeneous sample of children, only some of whom will have subsequent deficits in morphosyntax that are characteristic of SLI. Should there be wider (or different) causes of vocabulary delay than grammar delay and should diagnoses of delay weigh heavily on the ability that is being measured, then the resolution of delay may turn out to be a measurement artefact. In that case, the current approach of tracing developmental trajectories in a single model system would not be applicable.

Tables

Table s.1. Neurocomputational parameters that reliably discriminated between groups. TD = typically developing / no delay. PD = persistent delay. RD = resolving delay. Results are shown for two complementary statistical analyses. ANV = analysis of variance. Scores show partial eta-squared effect sizes. * = effect reliable at p<.05; ** = effect reliable at p<.01. MRL = multinomial logistic regression. Scores show Wald statistic for each parameter as a measure of effect size. Empty cells represent non-reliable differences (p>.05).

Parameter	Role	TD vs. PD		TD vs. RD		PD vs. RD	
		ANV	MLR	**ANV**	MLR	**ANV**	MLR
Hidden units	Capacity	** **.030**	* 6.6	* **.005**	* 4.6	** **.031**	
Architecture	Capacity	** **.018**	** 18.1	** **.013**	** 17.4		
Sparseness	Capacity						
Pruning onset	Capacity						
Pruning prob.	Capacity						
Pruning threshold	Capacity			* **.004**		* **.021**	* 5.4
Learning algorithm	Capacity / Plasticity	** **.172**	** 98.8	** **.012**	** 24.8	** **.104**	** 21.3
Learning rate (l-r)	Plasticity	** **.030**	** 16.4	** **.044**	** 33.2		
Semantic l-r	Plasticity	* **.005**				** **.024**	
Phonological l-r	Plasticity	** **.018**	** 7.0	** **.014**	** 8.6		
Momentum	Plasticity	* **.006**	* 4.7	** **.015**	** 12.8		
Weight variance	Plasticity			** **.009**	** 11.2		
Unit threshold function	Plasticity / Signal			** **.036**	** 23.0	** **.025**	* 5.5
Processing noise	Signal	** **.021**	** 19.1			** **.026**	
Response threshold	Signal	** **.038**	** 16.2	** **.063**	** 22.3		
Weight decay	Signal	** **.009**					
Fam. Quot. (SES)	Environment						

MLR model fit: TD versus all delay groups, X(72)=411.3, p<.001, Nagelkerke R^2=.405; persisting versus resolving delay group, X(18)=79.9, p<.001, Nagelkerke R^2=.328

Table s.2. Neurocomputational parameters that reliably discriminated between resolving delay groups. RDL = resolving delay with low outcome. RDG = resolving delay with good outcome. RDVG = resolving delay with very good outcome. Results are shown for two complementary statistical analyses. ANV = analysis of variance. Scores show partial eta-squared effect sizes. * = effect reliable at p<.05; ** = effect reliable at p<.01. MRL = multinomial logistic regression. Scores show Wald statistic for each parameter. Empty cells represent non-reliable differences (p>.05).

Parameter	Role	RDL vs. RDG		RDL vs. RGVG		RDG vs. RDVG	
		ANV	MLR	**ANV**	MLR	**ANV**	MLR
Hidden units	Capacity						
Architecture	Capacity						
Sparseness	Capacity						
Pruning onset	Capacity						
Pruning prob.	Capacity						
Pruning threshold	Capacity						
Learning algorithm	Capacity / Plasticity						
Learning rate (l-r)	Plasticity						
Semantic l-r **+**	Plasticity	**** .053**	* 6.1				
Phonological l-r	Plasticity						
Momentum	Plasticity						
Weight variance	Plasticity						
Unit threshold function	Plasticity / Signal	**** .064**	** 8.6				
Processing noise	Signal						
Response threshold	Signal						
Weight decay	Signal						
Fam. Quot. (SES)	Environment	**** .089**	** 11.9	**** .095**	** 12.6	**p=.058 .111**	

MLR model fit: comparison of three resolving groups, X(36)=57.3, p<.001, Nagelkerke R^2=.420; pairwise comparisons: RDL vs. RDG, X(18)=47.135, p<.001, Nagelkerke R^2=.417; RDL vs. RDVG, X(18)=4.844, p=.999, Nagelkerke R^2=.128; RDG vs. RDVG, X(18)=12.951, p=.794, Nagelkerke R^2=.567

+ Why did the semantic pathway learning rate discriminate between the RL and RG delay groups, and not the phonological pathway learning rate? This is due to the definition of delay according to regular verb performance. Verbs in the training set can be learned by a lexical strategy, facilitated by the lexical-semantic input. Generalisation of the past-tense rule is dependent on phonological similarity. Were delay defined according to regularisation of novel verbs, the phonological pathway learning rate would be the more salient parameter in modifying rate of development.

Table s.3. Mean values for neurocomputational parameters and environment for the typically developing / non-delay group (TD; N=713), the persisting delay group (PD, N=118), the resolving delay group with low outcome (RD-L, N=136), good outcome (RD-G, N=28), and very good outcome (RD-VG, N=5), as well as three individual case studies of persisting deficit (PD1), resolving low outcome (RD1), and resolving very good outcome (RD2).

Parameter	Role	Case studies			TD	PD	RD		
		PD1	RD1	RD2			L	G	VG
Hidden units	Capacity	22	20	22	31	22	27	29	22
Architecture	Capacity	0	1	1	1.08	.87	.93	.93	.80
Sparseness	Capacity	.2	.1	0	.06	.07	.06	.05	.10
Pruning onset	Capacity	100	100	50	105	100	104	99	130
Pruning prob.	Capacity	.1	.05	.5	.14	.15	.12	.10	.15
Pruning threshold	Capacity	.2	.75	.3	.53	.55	.51	.50	.42
Learning algorithm	Capacity / Plasticity	0	1	1	.97	.66	.91	.93	1.00
Learning rate (l-r)	Plasticity	.15	.075	.15	.13	.11	.11	.12	.13
Semantic l-r	Plasticity	.01	.1	.5	.55	.48	.55	.74	.65
Phonological l-r	Plasticity	.1	.5	.25	.38	.26	.28	.32	.32
Momentum	Plasticity	.6	.2	.1	.27	.23	.22	.23	.15
Weight variance	Plasticity	.5	1.0	.75	.53	.56	.64	.52	.70
Unit threshold function	Plasticity / Signal	.75	.75	.5	1.29	1.24	1.14	.82	1.00
Processing noise	Signal	.5	.5	1.0	.60	.83	.59	.63	.65
Response threshold	Signal	.1	.025	.025	.09	.04	.04	.03	.03
Weight decay ($x10^{-7}$)	Signal	9.80	0	0	5.37	15.70	3.57	7.07	4.52
Fam. Quot. (SES)	Environment	.96	.79	.96	.80	.79	.77	.86	.94

Figures

Figure s.1. Effect size of behavioural profile comparisons between individual whose outlier status on regular verbs persists from Time 1 to Time 5 compared to individuals who return to score in the normal range for regular verbs by Time 5. Effect sizes are shown for three behavioural metrics, regular verb performance (Reg), generalisation of the regular rule to novel stems (Rule), and performance on irregular verbs (Irreg); and for outliers defined by low initial performance (Delayed) versus those defined by high initial performance (Gifted), in each case scoring more than 1 standard deviation beyond the population mean.