Developmental Science 6:5 (2003), pp 537–556

PAPER

Multiple causality in developmental disorders: methodological implications from computational modelling

Michael S.C. Thomas

Neurocognitive Development Unit, Institute of Child Health, London

Abstract

When developmental disorders are defined on the basis of behavioural impairments alone, there is a risk that individuals with different underlying cognitive deficits will be grouped together on the basis that they happen to share a certain impairment. This phenomenon is labelled multiple causality. *In contrast, a developmental disorder generated by a single underlying cognitive deficit may nevertheless show variable patterns of impairments due to individual differences. Connectionist computational models of development are used to investigate whether there may be ways to distinguish disorder groups with a single underlying cause (homogeneous disorder groups) from disorder groups with multiple underlying causes (heterogeneous disorder groups) on the basis of behavioural measures alone. A heuristic is proposed to assess the underlying causal homogeneity of the disorder group based on the variability of different behavioural measures from the target domain. Heterogeneous disorder groups are likely to show smaller variability on the measure used to define the disorder than on subsequent behavioural measures, while homogeneous groups should show approximately equivalent variability. Homogeneous disorder groups should show reductions in the variability of behavioural measures over time, while heterogeneous groups may not. It is demonstrated how these predictions arise from computational assumptions, and their use is illustrated with reference to behavioural data on naming skills from two developmental disorder groups, Williams syndrome and children with Word Finding Difficulties.*

Introduction

Take a group of children with a particular developmental disorder, such as a problem with productive vocabulary. How do we establish what is atypical about the underlying cognitive system? The standard approach within psychology is to run behavioural experiments and look for anomalous but consistent patterns across individuals in the disorder group. Such patterns can indicate differences in the way representations or processes have developed. Perhaps these children demonstrate a particular delay in producing certain types of words, or a particular pattern of errors. Of course, individual differences and measurement error will add 'noise' to the data, but the hope is that consistent patterns characterizing the disorder will nevertheless emerge. Once the underlying cause of behavioural deficits is understood, we can begin to consider an appropriate form of remediation for the disorder, and use it theoretically to shed light on processes of normal development.

However, consider the following situation. What if there is no single cause for the behavioural impairment, if instead the behavioural metric used to define the disorder (such as slow production of words) can be caused by a number of different processing anomalies in the language system. In this case, the definitional measure will have brought together a disorder group with heterogeneous underlying causes. As a consequence, it may be difficult to discern a consistent pattern in the data generated from behavioural experiments.

The problem here is that 'noise' in experimental data may be either the result of individual differences (and measurement error) arising from a disorder group with a single underlying cognitive cause, or it may be the result of a disorder group in which there are multiple underlying causes, each producing its own consistent behavioural pattern (with individual variation and measurement error). Multiple patterns are then superimposed on each other when the results are viewed for the group as a whole. The question considered in this paper is, how can we tell the difference between these two possibilities? Presuming that we expect some degree of individual variation in people with a developmental disorder, what behavioural evidence will tell us whether a given disorder

Address for correspondence: Michael Thomas, School of Psychology, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK; e-mail: m.thomas@psychology.bbk.ac.uk

DESC_311 Pages: 20

group has a single underlying cognitive cause or multiple underlying causes? In this paper, this question is explored from the perspective of computational modelling, building on recent work in the modelling of atypical development and individual variation. First, however, this issue is briefly situated within the wider perspective of the study of developmental disorders.

If one excludes disorders caused by exposure to an impoverished environment, broadly speaking developmental disorders can be split into two groups: (1) disorders defined by a genetic aetiology (such as Fragile X syndrome, Down's syndrome, Williams syndrome, velocardiofacial syndrome, Turner's syndrome), and (2) those defined on behavioural grounds (such as autism, Attention Deficit Hyperactivity Disorder, Specific Language Impairment (SLI), and developmental versions of various acquired disorders such as dyslexia, dyscalculia, prosopagnosia and amnesia). Although precise genetics causes have not yet been identified for disorders within the behaviourally defined group, these disorders can often show significant degrees of heritability, for example as in the cases of SLI and autism (Bishop, North & Donlan, 1995; Pennington & Smith, 1997; Simonoff, Bolton & Rutter, 1998).

Where no evidence exists to identify robust sub-types of a disorder, it is assumed that each disorder can be characterized in terms of a distinct, atypically developing cognitive architecture. Attempts to identify this architecture then proceed via experimentation within a psychological framework. Such attempts must necessarily deal with the fact of individual variation: even when a disorder group can be identified as sharing an identical genetic anomaly (such as in the case of Williams syndrome), individuals can nevertheless show a large degree of variability in the behavioural deficits they exhibit (e.g. for Williams syndrome, see Pezzini, Vicari, Volterra, Milani & Ossella, 1999). Even when there is reason to believe that there is a single underlying cause at the cognitive level, this cause must be identified through a screen of individual variation. Variability is also to be expected in behaviourally defined developmental disorders, but now there is no independent basis to suggest a single underlying cause. Therefore, it is legitimate to ask whether the variability in the disorder stems from individual variation overlying a single cognitive cause, or from different underlying cognitive causes, each with its own individual variability. Such debates can already be found in the study of Specific Language Impairment (e.g. Tomblin & Pandich, 1999; van der Lely, 1999) and in the study of developmental dyslexia (e.g. Fletcher, Foorman, Shaywitz & Shaywitz, 1999). Importantly, the fact that many behaviourally defined developmental disorders are associated with high levels of heritability does not preclude the possibility that they have multiple underlying causes, since *each* of the multiple causes could itself produce an equivalent level of heritability. High heritability does not necessarily imply single underlying cause.

Even though *multiple causality* in behaviourally defined disorders is a distinct possibility, the difficulty lies in how we identify that it is present in a given disorder group. Specifically, we currently lack a behavioural marker for multiple causality because we cannot systematically compare what sort of data would be produced by a disorder group with a single underlying cause and one with multiple causes. Computational modelling, on the other hand, provides a controlled, formal environment in which exactly this sort of question can be considered. It provides the opportunity to construct models which have developmental deficits, and to define in advance disorder groups that share a common underlying deficit versus those that contain individuals with different underlying deficits. By comparing these groups, we may then seek to identify a behavioural marker of multiple causality. First, however, it is important to understand how computational models have been used to study atypical development and individual differences.

Computational models of typical and atypical development

The increasing emergence of computational modelling as an approach to the investigation of typical development has made possible an extension of these techniques into the realm of atypical development. Much of the modelling of typical development has taken place within the connectionist framework (see e.g. Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996). It is this framework that will form the focus of this article.

A connectionist model of typical development begins by formulating the chosen cognitive domain as a set of inputs or input-output mappings. These are then encoded in a psychologically plausible representational format. The training set is applied (according to some regime) to a connectionist network deemed appropriate for acquisition of the cognitive domain. A successful model will exhibit an improvement in performance that follows the developmental trajectory found in children, as well as the final level of competence found at the end of development. Implicit in such models are computational constraints built into the system prior to the developmental process. These constraints shape the subsequent developmental trajectory when the training set is applied to the model. The constraints include the initial architecture, the representational scheme used to depict the cognitive domain as a set of mappings, the learning rule and the dynamics of activation flow through the network.

Connectionist models of developmental disorders assume that differences in these initial computational constraints are the cause of subsequent atypical trajectories of development, trajectories which may include behavioural impairments during or at the end of the developmental process (see Thomas & Karmiloff-Smith, 2002a, for a review). For example, categorization deficits in *autism* have been modelled in a system which possesses an atypical initial architecture (Cohen, 1994, 1998). *Developmental dyslexia* has been modelled by various researchers using models which possess atypicalities in initial architecture, in initial activation dynamics or in the initial learning rule (e.g. Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989). Deficits in inflectional morphology in *SLI* have been modelled in a system with initial atypicalities in representational schemes (Hoeffner & McClelland, 1993), and deficits in inflectional morphology in *Williams syndrome* have been investigated in a model experiencing a variety of changes in initial constraints to the processing of semantic or phonological knowledge (Thomas & Karmiloff-Smith, 2002b).

This approach to explaining deficits in developmental disorders accords with the recently elucidated *neuroconstructivist* approach (Elman *et al.*, 1996; Karmiloff-Smith, 1998; Oliver, Johnson, Karmiloff-Smith & Pennington, 2000), which argues that despite behavioural similarities, no strong analogies can be drawn between developmental deficits and acquired deficits which emerge in adults who have experienced brain damage. While the latter may well correspond to selective damage to high-level cognitive modules, the former will be the consequence of atypical trajectories of development operating under different low-level neurocomputational constraints. In the case of genetic disorders, it is likely that genetic anomalies will act on these low-level neurocomputational constraints, rather than coding directly for (the success or failure of) high-level cognitive modules. In the developmental disorder, the developmental process itself is a key mediator between genome and phenotype (Karmiloff-Smith, 1998; see Thomas & Karmiloff-Smith, in press a, for a detailed computational consideration of the relation of acquired and developmental deficits). The advantage of computational models in this context is that they move away from a static consideration of developmental deficits, and allow for a much more precise consideration of dynamic explanations of developmental deficits that focus on the developmental process itself as a key causal element in generating the disorder.

Multiple causality in models of disorders

The computational modelling of developmental dyslexia illustrates one important idea to emerge from this work on atypical development. It is that a given behavioural impairment (such as poor reading of exceptions words in dyslexia) can be generated by more than one atypical computational constraint. For example, poor exception word reading was simulated in separate studies by reducing the number of internal processing resources available in the model, by slowing down the learning rate, and by using an inefficient learning algorithm (Bullinaria, 1997; Harm & Seidenberg, 1999; Plaut, McClelland, Seidenberg & Patterson, 1996; Seidenberg & McClelland, 1989; see Thomas & Karmiloff-Smith, in press a, for review). Each of these manipulations punishes the learning of patterns that are less strongly encoded in the training set.

Thomas and Karmiloff-Smith (2002b) carried out a more systematic exploration of the potential for multiple causality in such models. This work initially set out to test computationally the viability of six competing hypotheses for explaining the pattern of performance shown by individuals with Williams syndrome in one area of language development, the acquisition of the English past tense. Starting with a model of typical development in the domain, initial computational constraints were varied in line with each theoretical hypothesis, and the subsequent atypical trajectories compared against empirical data. Importantly, Thomas and Karmiloff-Smith did not stop with the evaluation of just these six hypotheses; they also went on to explore the background flexibility of the model in producing different patterns of impairments under a range of changes to the initial computational constraints in the model. The results demonstrated two points. First, for a given behavioural measure, there were often several initial constraints that would generate a deficit in this measure. For example, generalization of regularities within the past tense domain to novel exemplars (such as the 'add -ed' rule) could be impaired by initial changes to the architecture (use of a four-layer rather than three-layer network; use of fewer hidden units), by initial changes to the representational schemes (a reduction in the similarity between training exemplars) or by initial changes to processing dynamics (processing noise, decreased discriminability in the activation function of units). Second, changes in initial computational constraints did tend to produce distinct *patterns* of deficits (and sometimes improvements) across *sets* of behavioural measures pertaining to the domain, in this case six measures used to assess the overall performance of the system.

These simulations confirm that, to the extent that connectionist systems are valid models of development, multiple causality of deficits is highly plausible. They also indicate that causes of deficits are only ambiguous when the deficits are narrowly defined on a restricted number of measures. The wider the range of measures used to assess a system, the smaller the chance of multiple causality, since a *pattern* of deficits across several measures is likely to have a unique cause. The narrower the deficit used to define a behavioural disorder, the greater the chance of recruiting systems into the disorder group which have different computational causes underlying their deficit and which overlap only on this particular deficit.

We have, then, a computational framework in which issues of single or multiple causality in the genesis of developmental deficits can be considered. But for the purposes of this paper, a computational theory of individual differences is also required, since it is individual variation that intervenes between the cognitive cause(s) of a disorder and our ability to assess it empirically through group studies.

Computational accounts of individual variation

Little systematic work on individual variation has been carried out in connectionist psychology. When Thomas and Karmiloff-Smith (in press b) reviewed this area, they found that researchers had proposed that individual variation could be accounted for by manipulations to the same computational parameters previously employed to capture other forms of cognitive variation. For example, an alteration to the level of internal processing resources (numbers of hidden units) in connectionist networks has been proposed as a candidate to explain individual variation, but also (independently) as a candidate to explain change in performance across cognitive development, and as a candidate to explain changes in ageing, and as a candidate to explain atypical trajectories of development. The proposal that *all* forms of cognitive variation can correspond to changes to the same underlying parameter is unlikely to be theoretically tenable. However, it is perhaps unfair to consider this *de facto* position as an explicit theory, given the very recent application of cognitive modelling across the full set of domains. A more coherent and differentiated position may emerge with time.

With respect to connectionist accounts of individual differences, two things can currently be concluded. First, many connectionist researchers (sometimes implicitly) take the view that differences in the initial random connection strengths in their networks and differences in the random order of exposure to items in the training set together form a possible explanation of individual variability. Second, this view is almost certainly wrong, both on neurobiological grounds (i.e. it seems unlikely that normal brains are anatomically identical save for initial differences in connection strengths) and on empirical grounds (i.e. to explain the general factor of intelligence by appeal to random initial connectivity alone, one would have to make the speculative leap that certain individuals have beneficial initial random weights in all the separate components of their cognitive systems). If we must appeal to other constraints to explain individual variation, it then becomes necessary to tackle the question of whether these constraints will be the same as those that explain cognitive development or atypical development or ageing. There are many possibilities. It may be, for example, that atypical lies at the extreme end of a continuum of individual variation, but that cognitive development and ageing correspond to quite different computational parameters. A discussion of these issues can be found in Thomas and Karmiloff-Smith (in press b).

Despite the preceding comments, in the following simulations, the assumption was indeed made that individual differences can be generated by differences in initial random weights and in training regime! Despite the wider problems associated with this assumption, for modelling purposes it permitted the definition of a source of computational variability that could be applied to typically developing and atypically developing groups alike. The variation of random initial weights and training regime within a small range was used to simulate individual differences, while variation of other initial computational constraints to more extreme values was used to create disordered development of various types. The important assumption here is not about the source of individual differences *per se* (one could shift the assignment of some of the computational constraints from the atypical set to the individual differences set, or define ranges of variation that correspond to individual differences versus those that correspond to atypical development). Rather, the important assumption for the simulations is that the variation of one (or more) constraints within a certain range explains individual variability *across all groups*, while the variation of the same constraint or other constraints to more extreme values explains the divergence of atypical from typical development.

The simulations

The following simulations are, to my knowledge, the first to address the implications of individual variability in behaviourally defined developmental disorders. Their aim was to generate heuristics that could separate behaviourally defined disorder groups of two types: those with a single (homogeneous) underlying cause versus those with multiple (heterogeneous) underlying causes. The advantage that computational models conferred here was that *it was known in advance what was wrong with each network*, because certain initial computational constraints had been deliberately manipulated. It was therefore possible to create a disorder group containing individuals that shared a single cause and a group that was a collection of several causes; and subsequently, to step outside the model and examine whether the two groups could be distinguished merely on behavioural grounds. Moreover, the full developmental trajectories of each individual network were accessible, allowing comparison of groups not only at the 'point of diagnosis', but also longitudinally. It could therefore be determined whether, even if homogeneous and heterogeneous groups were initially indistinguishable, they nevertheless diverged over developmental time.

In the following simulations, then, the procedure was as follows: (1) for a given domain, define a developmental disorder according to a specific behavioural impairment; (2) establish two groups of networks showing the impairment, one comprising individual networks with a homogeneous underlying computational cause, one comprising networks with heterogeneous underlying causes; (3) search for behavioural grounds that would allow the two groups to be distinguished, either at a single point in developmental time or longitudinally.

The model

Below, a qualitative description of the design and results of the simulation work is given. Technical details can be found in Appendix A. The simulations were based on the well-explored developmental domain of past tense formation, and employed a standard three-layer backpropagation network and a training set adapted from Plunkett and Marchman (1991, 1993).

The simulations can be viewed in abstract terms, as corresponding to a notional cognitive system required to learn a set of input-output mappings in a domain characterized by a partial regularity (i.e. where the majority of the mappings conform to a rule but a minority form exceptions). On the other hand, they can be viewed in much more concrete terms, as corresponding to the simulation of the English past tense acquisition. A number of existing developmental disorders have been evaluated with respect to this domain, including SLI (Montgomery & Leonard, 1998; Moore & Johnston, 1993; Oetting & Horohov, 1997; Ullman & Gopnik, 1999; van der Lely & Ullman, 2001), Williams syndrome (Bromberg, Ullman, Coppola, Marcus, Kelley & Levine, 1994; Clahsen & Almazan, 1998; Thomas, Grant, Gsödl, Laing, Barham, Lakusta, Tyler, Grice, Paterson & Karmiloff-Smith, 2001), spinal muscular atrophy (Sieratzki & Woll, 1998), early and continuously treated phenylketonuria (Badali, Izvorski, Ozawa, Diamond & Ullman, 1999), and children with non-specific developmental disabilities (Newfield & Schlanger, 1968). This aspect of language development has been of great theoretical interest since children must acquire a rule-based inflectional paradigm (English verbs form their past tense by adding -ed to the verb stem) in the face of exceptions to these rules (e.g. sleep-slept, hit-hit, go-went). This is a task that some researchers argue requires qualitatively different cognitive mechanisms to reflect the two types of mapping. This debate is not directly relevant to our goals here, and no strong claims are made that this is an ideal model of past tense disorders. Conceiving of the model in relation to this domain does, however, provide a more tangible grounding for the simulation data to be encountered.

In abstract terms, the model domain corresponds to the acquisition of regular and exception patterns. Here, then, are two initial behavioural metrics for the domain: how well does the system acquire regular mappings, and how well does it acquire exception mappings? A third metric can be defined according to error patterns, specifically the prevalence of a particular class of errors where the regularity in the training set is mistakenly overextended to the exception mappings. The way the system generalizes its knowledge to novel exemplars provides three more metrics. These are, first, in terms of the extension of the regularity to novel exemplars which are similar to existing regular mappings; second, in terms of the extension of the regularity to exemplars which are similar to existing exception mappings; and third, in terms of generalization of exception patterns to novel exemplars sharing a similarity with existing exception mappings. Together, these six metrics allowed *patterns of deficits* to be probed for in the disordered networks.

In terms of past tense formation, the model can be viewed as learning to associate phonological representations of verb stems with their corresponding past tense forms. The six metrics of performance can then be rephrased as follows: (1) levels of acquisition of the regular past tense (e.g. *talk-talked*); (2) levels of acquisition of the set of irregular past tenses (e.g. *think-thought*, *sleep-slept*); (3) rates of over-generalization errors on irregular past tenses (e.g. *thinked*, *sleeped*); (4) generalization of the regular past tense to novel verbs dissimilar to existing exception verbs, which will be termed 'non-rhymes' (e.g. *vask-vasked*); (5) generalization of the regular past tense to novel verbs rhyming with existing exception verbs, termed 'rhymes' (e.g. *crive* rhymes with *drive*, *frink* rhymes with *drink*, past tense generalized as *crived* and *frinked*); and (6) generalization of irregular past tense patterns to these 'rhymes', termed 'irregularization' (e.g. *crive*-*crove*; *frink*-*frank*).

In Study 1, a developmental disorder was defined on the following behavioural grounds. Partway through training, certain individual networks were found to show exaggerated levels of the error involving over-extension of the regularity to exception patterns (i.e. *over-generalization*). The networks were approached as if they suffered from a mysterious developmental disorder marked by over-generalization, and designed a notional empirical study to investigate what might be the underlying cause of this disorder. The study compared the behaviour of the disordered networks with that of normally developing networks of the same 'chronological age' (i.e. evaluated at the same point in training) to establish the level of impairment. It then compared the disordered networks to a set of younger, less-trained networks. These younger networks also exhibited higher levels of the over-generalization error, but in this case as a sign of immaturity. The second comparison allowed the investigation of whether the mystery networks were 'simply showing symptoms of developmental delay', a standard hypothesis in the field. The comparison of individuals with developmental disorders against chronological-age-matched controls and against mental-age-matched controls is a common empirical paradigm in the study of atypical development. Lastly in this notional study, individual networks were also examined longitudinally to establish their final performance outcomes.

Elevated levels of over-generalization errors in past tense formation have been reported in several studies of developmental disorders, including children with WS, spinal muscular atrophy, and early and continuously treated phenylketonuria. In the current case, however, the 'over-generalization' disorder is solely viewed as a behavioural disorder. (Such a criterion might seem rather narrow, but this is not unprecedented. For instance, the primary behavioural definition of developmental phonological dyslexia is a difficulty in naming individually presented novel letter strings [see Fletcher *et al.*, 1999].) This notional study of the past tense disorder can be roughly calibrated against existing empirical data (e.g. Thomas *et al.*, 2001). The study corresponds to children who exhibit the 'over-generalization' disorder at age 10. They are compared to typically developing children of the same age (the chronological-age match) and to younger children of about 6 years of age (the mental-age match), where in all cases, the individual children are followed longitudinally to evaluate their final past tense performance as adults.

Suffice to say, the claims made about multiple causality are intended to be more general than the past tense domain, for reasons discussed later. Indeed one of the resulting heuristics will be illustrated via empirical data from an unrelated domain. The precise definitions of the disorder and control groups will now be considered.

Study 1

A behavioural developmental disorder was defined based on the demonstration of elevated level of over-generalization (OG) errors occurring midway through training, against a baseline of typically developing networks. Several atypical initial computational constraints could produce a developmental trajectory exhibiting such elevated errors. There is insufficient space here to discuss the computational reasons why each manipulation has the developmental effect it has – readers are directed to Thomas and Karmiloff-Smith (2002b). However, these reasons are not central to the argument. Four test groups were established, each with ten individual networks trained with different initial random weights and order of presentation of the training set. The groups were as follows:

1. *A homogeneous disorder group*. At the designated point in training (500 exposures to the training set), this group exhibited a mean level of OG errors of 35.9% of exception patterns, approximately ten times higher than the normal rate. A single computational cause was responsible for the impairment in this disorder group, specifically the use of localist rather than distributed representations to represent the phonology of the verb stem and past tense (see Appendix A). These networks were a random set of networks trained with this initial impairment.
2. *A heterogeneous disorder group*. At the designated point in training, this group of ten individuals exhibited a mean level of OG errors of 34.4% of exception patterns. Five different computational causes were responsible for this impairment in the heterogeneous group, with two individuals suffering from each 'condition'. Two individuals shared the localist impairment of the homogeneous group (but were different individuals); two individuals used a compressed representational scheme that while still distributed, exaggerated the similarity between phonemes; two individuals were trained with a slower learning rate; two individuals were trained with a learning algorithm with reduced plasticity; and two individuals had a network architecture without a hidden layer. Other computational constraints may also cause

increased OG errors: the particular conditions used were chosen at random from those producing individual networks with meeting the behavioural criterion. Note that not every individual network suffering from these five separate atypical conditions necessarily generated the same high level of OG errors: the individual networks used in the heterogeneous group were hand picked to be precisely those with a combination of atypical initial constraints and elevated OG errors. But it must be emphasized, *this is the whole point of a behaviourally defined developmental disorder*: the individuals recruited may as easily be outliers from one distribution of individuals with a particular deficit as central members of another distribution. The criterion picks the individuals, not the underlying cause.

3. *A 'chronological age' (CA) matched control group*. At the designated point in training, this typically developing control group exhibited a mean level of OG errors of only 2.9% of exception patterns. These networks were a random set of networks trained with normal initial constraints.
4. *A 'mental-age' (MA) matched control group*. For the individual networks in the CA control group, performance was traced back to an earlier point in training, when OG errors were more frequent. After just 100 presentations of the training set, the error level was approximately equivalent to the disorder groups, at 32.8% of the exception patterns. This group of 'younger' individuals was then defined as the MA control.

T-tests revealed no significant differences between the MA control group and both disorder groups on the metric used to define the disorder, OG errors (all $p > .2$). On this measure alone, neither disorder group was distinguishable from delayed development. Subsequently, performance was compared across the other five performance metrics, and longitudinally at the end of training (5000 exposures to the training set).

Study 2

For reasons that will become apparent shortly, it was necessary to validate the initial results by defining a second behavioural developmental disorder. This was based on the criterial measure of low performance on exception patterns (which did not turn out to have a directly complementary relationship to the level of OG errors since other types of error were possible). For this 'poor exception' disorder, CA, MA groups were the same, a new random set of individuals was generated for the homogeneous disorder group, and a new heterogeneous disorder group was recruited to give a closer match to the homogeneous group on the new metric. The new heterogeneous group contained none of the individual networks from Study 1.

1. *Homogeneous disorder group*. A new set of ten networks trained with different initial randomized connection weights, and with the 'localist' impairment.
2. *Heterogeneous disorder group*. This group comprised ten individual networks with six different underlying computational causes of their disorder. One individual employed the compressed representational scheme; two individuals employed processing units with reduced discriminability; two individuals employed the learning algorithm with reduced plasticity; three individuals used networks in which the initial weights were randomized with greater initial variance;[1] one individual used a network with no hidden layer; and one individual used a network in which the number of units in the hidden layer was reduced.
3. *CA-matched control group*. As for Study 1.
4. *MA-matched control group*. As for Study 1.

For the CA group, performance on exception patterns was at 85.6% after 500 exposures to the training set. For the homogeneous disorder group, performance was at 20.3%, while that for the heterogeneous group was at 20.6%. The younger MA networks performed at 25.7%. Homogeneous and heterogeneous groups did not differ significantly from one another ($p > .8$) on the definitional measure. The MA group was not as well matched for this study, exhibiting slightly better performance than both disorder groups ($p = .005$ and $p = .047$ for the homogeneous and heterogeneous groups, respectively). Once again, performance was examined across behavioural metrics other than that used to define the disorder, as well as longitudinally for all individual networks.

Results and discussion

Multiple causality

Employing simple connectionist learning models of the type widely used in the study of cognitive development, it is relatively straightforward to demonstrate that a behavioural developmental disorder can have multiple underlying computational causes. The disorder defined in Study 1 had five potential causes and the disorder in Study 2 had six potential causes, yet in both cases, the

[1] This is an example where atypical development and individual differences are generated by the same computational constraint. Initial weight randomization within a small range contributes to individual differences. Set to an extreme value, it generates atypical development.

homogeneous and heterogeneous disorder groups were not significantly different on the criterial measure for the disorder. Note, however, that multiple causality is not a question of *computational equivalence* between the underlying causes. Particular computational causes tend to produce a unique pattern of changes across a set of performance metrics, even though these causes often overlap on individual metrics (see Thomas & Karmiloff-Smith, 2002b). The narrower the defining metric for a behavioural developmental disorder, the more likely it is that the disorder will have multiple underlying causes. Disorders defined across wider patterns are more likely to have single underlying causes.

The point here is perhaps a slightly subtle one, because it might be argued that a narrowly defined behavioural disorder will certainly recruit a more similar set of individuals than a *vaguely* defined disorder, which sweeps together individuals who really don't have the same type of disorder. For example, the definition of phonological developmental dyslexia (problems with decoding single novel words) is much narrower than 'any child who has a problem with reading' and surely more likely to unify children with a similar sort of deficit. The argument here distinguishes between two different alternatives to narrow: narrow vs. vague, and narrow vs. wider-but-clearly-specified. Thus a narrowly defined disorder will recruit a more homogeneous group than a vaguely defined disorder. But a disorder that is strictly defined across a greater number of behavioural measures relevant to the target domain is likely to increase the homogeneity of the group still further, and increase the chances of indexing a unique underlying computational cause.

Do additional behavioural measures per se *help to differentiate disorder groups?*

It may seem obvious that if one collects more behavioural measures, one is automatically more likely to gain a greater basis to distinguish homogeneous and heterogeneous disorder groups. The results show that this *can* be the case. In Study 2, although the homogeneous and heterogeneous disorder groups did not significantly differ in their mean performance on the defining behavioural metric, they showed differences significant at the .05 level in four of the other five behavioural metrics. However, it is not *necessarily* the case. In Study 1, again homogeneous and heterogeneous disorder groups did not significantly differ on the definitional metric, but *nor did they differ* on mean performance in four of the other five performance metrics. On only one metric (regularization of novel non-rhymes) could the groups be distinguished on their means ($p < .001$; all other $p > .3$). Extra measures do not necessarily help.

This result arises because, as will become apparent, the heterogeneous group demonstrated larger variability in its individual scores, and this variability was sufficient to eliminate many significant differences between the disorder groups. (As an illustration, the MA controls can be used as a neutral baseline. Compared to this baseline, the homogeneous group was distinguishable on the mean values of four of the five additional measures, while the heterogeneous group was not distinguishable from MA controls on any of the further measures.)

Parametric tests such as the *T*-test assume homogeneity of variance between the comparison groups. Therefore, one possibility why a heterogeneous disorder group can be statistically indistinguishable from a homogeneous disorder group on mean scores is that the variability of the two groups is not equivalent, violating the assumptions of the parametric test. However, this appears not to be the explanation here: use of the non-parametric Mann-Whitney U test produced the same pattern of results. The heterogeneous disorder group was indistinguishable from homogeneous and MA control groups due to greater overlap of scores with other comparison groups, rather than greater score variability *per se*.

In short, there is no automatic guarantee that by collecting more performance metrics relevant to the domain and comparing *mean scores*, homogeneous and heterogeneous disorder groups can be disambiguated on behavioural grounds.

Will time always differentiate disorder groups?

One might also expect that, even if the two disorder groups are indistinguishable at a certain point in development, time will subsequently separate them. When individual performance is examined longitudinally, surely we should expect the homogeneous and heterogeneous disorder groups to diverge in terms of their mean scores? Again, the answer is, not necessarily so. In Study 1, at the end of training, the two disorder groups were not significantly different at the .05 level on *any* of the six performance metrics. In Study 2, the groups differed on only one of the six measures (interestingly, this was not even the original criterial measure for the disorder). The same pattern was revealed using the non-parametric Mann-Whitney test. Again, the explanation here lies in terms of variability and overlap. Although networks with different computational causes did diverge in their performance over development, the consequence was greater variability in the heterogeneous group, increasing the overlap between the groups, and denying the ability to distinguish significant differences between the mean levels of performance at the end of training.

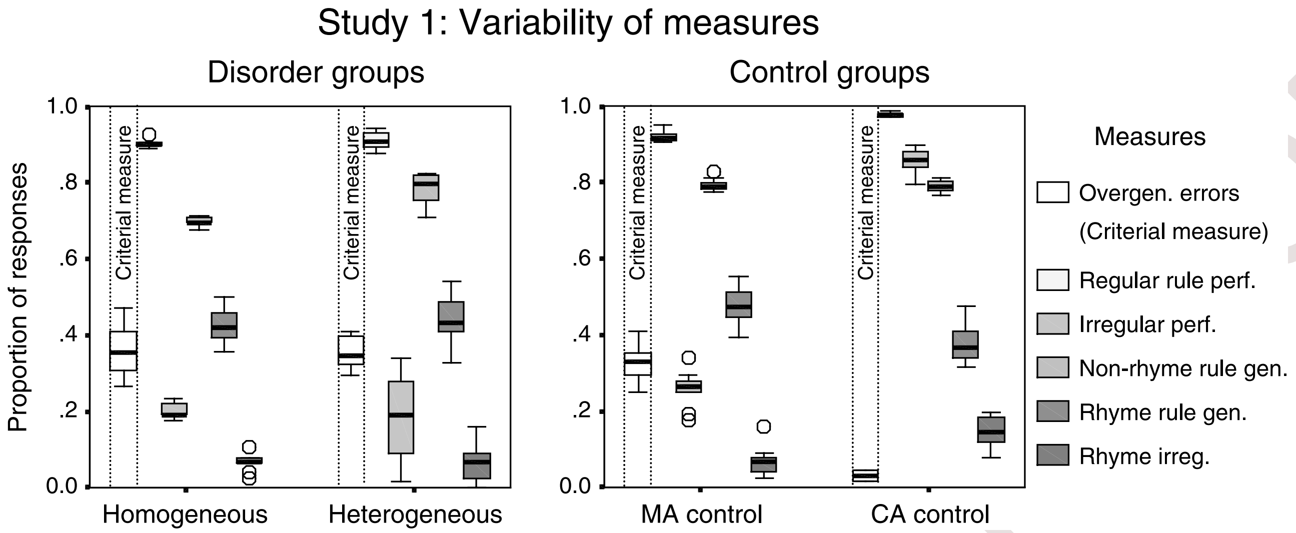

Figure 1 *Boxplots showing the variation in scores produced by individual networks in the two disorder groups and the two control groups in Study 1 (N = 10 per group). For each group of individuals, the left-most box (white) indicates variability in the criterial measure that defines the presence of the disorder. To the right of this measure, variability is then shown for the five other metrics from the cognitive domain. (For each box, the central line represents the median value of the group; the box captures the middle 50% of the cases; the whiskers connect the largest and smallest values that are not categorized as outliers or extreme values; 'o' represents an outlier more than 1.5 box-lengths away from the box; '*' represents an extreme value more than 3 box-lengths away from the box.)*

The use of variability to distinguish disorder groups

The news is bad thus far: in behavioural terms, the homogeneous and heterogeneous disorder groups were not necessarily distinguishable by examining multiple measures, nor by examining performance longitudinally. However, things look much more optimistic when we move beyond mean scores to examine the *variability* of scores. Here, the results suggested heuristics that may be successful in distinguishing homogeneous and heterogeneous groups. Figure 1 contains boxplots depicting the variability for all test groups in Study 1.

Two points are of note. The first is that the heterogeneous disorder group in fact showed lower variability than the homogeneous group on the measure that defines the disorder (left panel; a standard deviation of 4.2% against 5.9%). This is less surprising when one recalls that individuals in the heterogeneous group were specifically recruited according to this measure. Second, it tended to be the case that for the homogeneous disorder group, the variability of the five non-definitional measures was smaller than the variability of the definitional measure. On the other hand, for the heterogeneous group, the variability of the non-definitional measures was larger or no different. In Figure 1, for the homogeneous group, all five additional measures showed smaller variability than the criterial measure; in four cases Levene's test for homogeneity of variance (Levene, 1960) suggested the difference was significant. For the heterogeneous group, three additional measures exhibited larger variance. Levene's test suggested one additional measure had significantly larger variability, one smaller, and three not significantly different (values are shown in Table 1).

Why does the change in variability across additional measures differ between homogeneous and heterogeneous groups? As suggested before, atypical computational constraints in developmental models tend to result in a pattern of changes across performance metrics, according to the particular limitations that these constraints place on the development of internal representations. For a given set of constraints, the pattern should be similar for each network despite blurring by individual differences. The variability for each measure within the pattern should thus be broadly similar, ceiling and floor effects aside. On the other hand, in the heterogeneous group, there are several different patterns unified only by their intersection on a single measure. It is therefore unsurprising that the variability of scores can be larger on the other measures where the patterns do not coincide.

Of course, inspection of Figure 1 suggests that the base rate variabilities of the six measures themselves vary even in the control groups. This stems from the problem domain, and the solutions that network systems

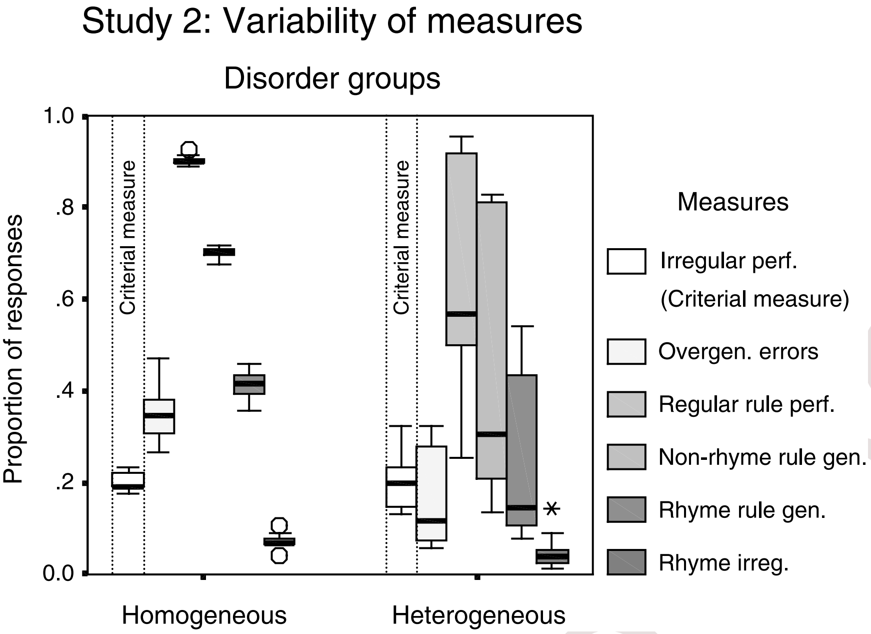

Figure 2 *Boxplots showing the variability of measures for the homogeneous and heterogeneous disorder groups in Study 2.*

can find to acquire the domain given their structure. Specifically, the higher net frequency of regular mappings in the training set strongly drives learning and forces the networks to dedicate most of their internal representational resources to reducing error on these mappings. Given that the representations are shaped by regular patterns, exception patterns must then be fitted 'around the edges', and it turns out that there are more solutions to this problem than there are to learning the regular verbs, and more ways to go wrong. The consequence is greater variability in irregular performance than regular performance.[2]

In Study 1, by chance the measure chosen to define the disorder happened to be the one that showed the largest variability in the homogeneous disorder group. It was possible, therefore, that the subsequent finding of smaller variability on the other measures in this group was artefactual. To check for this possibility, the exercise was repeated in Study 2, but now choosing a definitional measure on which the homogeneous disorder group had a smaller variability. A new heterogeneous group was then recruited according to this new definitional measure. The comparison of variability for these two groups can be seen in Figure 2. Although by design it was now true that the homogeneous group demonstrated larger variability across some of the additional measures, comparison with the new heterogeneous group reveals that the increase in variability across additional measures was very much greater for the heterogeneous group. Here for the homogeneous group, one additional measure had statistically greater variance than the criterial measure, one had less variance and three were not significantly different. For the heterogeneous group, four additional measures had significantly greater variance, while one was not significantly different (values shown in Table 1).

Distinguishing the disorder groups using variability over time

Finally, the effect of developmental time on the variability of measures was examined. Figures 3(a) to 3(c) depict respectively the variability of measures for the control group, the homogeneous disorder group and the heterogeneous disorder group in Study 1 at two points in time; first, the initial stage when the disorder was 'diagnosed', and second, the longitudinal outcome at the end of training.

[2] For partially regular domains of this sort, it turns out to be hard to specifically disrupt the learning of regular patterns. Generally, changes in initial constraints cannot significantly deflect the trajectory of regular learning without impairing all learning in the network. Hoeffner and McClelland (1993) managed to specifically target regular mappings in acquiring the past tense domain by altering the initial representational scheme to weaken the information that encoded the regularity, effectively redefining the nature of the domain. Marchman (1993) managed to produce a model in which learning of regulars was weaker than learning of irregulars, by increasing the frequency of the irregular mappings such that they were able to dominate the regular mappings in determining the shape of the internal representations.

Table 1 *Between-measure comparisons of variability*

	Measure: Group	*1(Criterial)*	2	3	4	5	6
Study 1	Homogeneous	*.062*	.010*	.021*	.011*	.045	.021*
	Heterogeneous	*.045*	.022*	.115*	.042	.061	.056
Study 2	Homogeneous	*.021*	.059*	.010*	.012	.032	.018
	Heterogeneous	*.060*	.107*	.252*	.307*	.188*	.039
Control	MA	*.046*	.016*	.047	.015*	.049	.038
	CA	*.012*	.005	.031*	.013	.051*	.043*

Note. Scores show standard deviations for Criterial measure (italic) and Additional measures (* = within-group difference between Criterial measure and Additional measure significant at 0.05 level, using the Levene test for homogeneity of variances (Levene, 1960)).

For the control group, the variance of five of the six measures reduced over time (using Levene's test, 3/6 significant reductions, 3 no change; values shown in Table 2). The homogeneous disorder group also demonstrated reductions in the variability of several of its measures over time (2/6 significant reductions, 4 no change). The reductions occurred because each network was approaching the best performance it could manage given the initial computational constraints it inherited. By the end of training, initial differences in weights and in the order of presentation of training items played a lesser role compared to the fit of the network constraints to the structure of the training problem. However, Figure 4 demonstrates that this reduction in variability was much less apparent in the heterogeneous disorder group. Variability remained large (5/6 no change, 1 significant increase in variability). This is because the performance ceilings that the individual networks in this group approached were different. Therefore, no convergence occurred and no reduction in variability. Figure 6 compares the endstate variability for the homogeneous and heterogeneous disorder groups in Study 2. Again, reductions in variance occurred for the homogeneous group (2/6 decrease, 4 no change) but were less marked for the heterogeneous group (1 significant increase, 3 no change, 2 significant decreases).

Summary

These simulations have suggested three ideas: (1) that multiple causality is a problem when the definition of a behavioural disorder is narrow with respect to the number of measures it incorporates; (2) that simply collecting more behavioural measures pertaining to the target domain will not necessarily distinguish disorder groups with homogeneous and heterogeneous causes, so long as *mean performance levels* alone are compared; but (3) that these groups can be distinguished on the basis of *variability*, either by comparing the variability of new measures against that of the defining measure, or by observing the variability of measures in each group over time.

An empirical example

The first heuristic – variability of additional measures – can be illustrated with regard to data from an unrelated domain, naming difficulties. Children who experience problems with productive vocabulary are described as having Word Finding Difficulties (WFD), and may be diagnosed on behavioural grounds using standardized tests. Children can exhibit naming difficulties while having non-verbal intelligence scores in the normal range, and exhibiting no major difficulties in articulation (Dockrell, Messer & George, 2001). Alternatively, naming difficulties can be found in other disorders with known genetic aetiologies, such as Williams syndrome (e.g. Temple, Almazan & Sherwood, in press; Thomas, Dockrell, Messer, Parmigiani, Ansari & Karmiloff-Smith, 2002). Dockrell and colleagues have suggested that WFD may well be a behavioural disorder with heterogeneous underlying causes within the language system (Dockrell *et al.*, 2001). On the other hand, with no evidence currently demonstrating robust sub-groups in Williams syndrome (WS), this genetic disorder is currently viewed as having a single underlying cause at the cognitive level. Using a sample of individuals with WS and a sample of individuals with WFD, naming therefore provides a domain where we may compare issues of cross-measure variability in a homogeneous and a (possibly) heterogeneous developmental disorder group.

Word Finding Difficulties in children are diagnosed initially by speech and language therapists based on evidence of difficulties in word finding in spontaneous speech and patterns of substitution errors. Subsequently a specific behavioural test (the Test of Word Finding Difficulties; German, 1989) assesses naming accuracy in response to semantic cues, revealing where there are unexpected production difficulties given the child's level of language comprehension. Unfortunately, no data are available for this test for individuals with WS. Instead, comparable data are available for a standardized test of productive word fluency (PhAB; Fredrickson, Frith & Reason, 1997). This test assesses how many words a participant can produce from a certain category (semantic

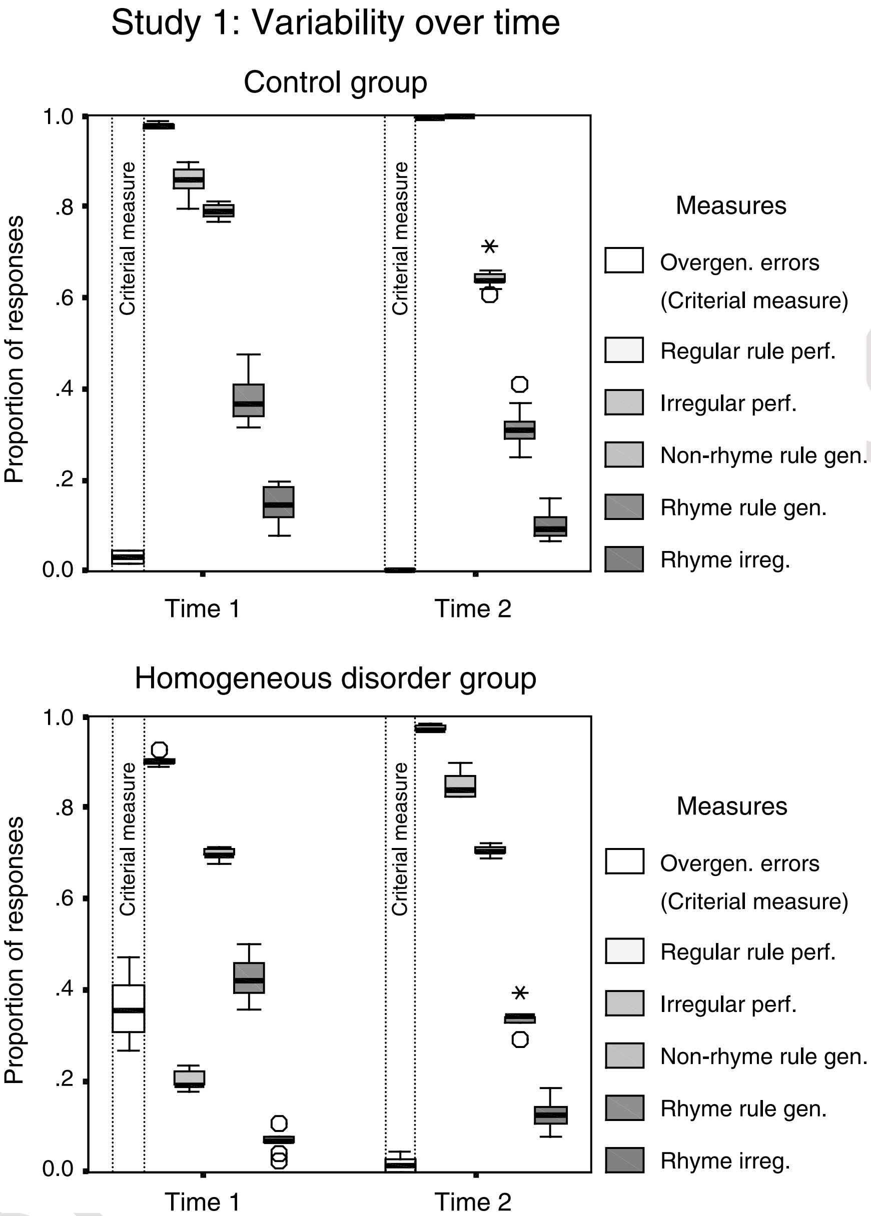

Figure 3 *Boxplots depicting the variability of the measures over time for (a) the control group; (b) the homogeneous disorder group; and (c) the heterogeneous disorder group in Study 1. Time 1 corresponds to the point of 'diagnosis' of the disorder (500 presentations of the training set). Time 2 corresponds to the end of training (5000 presentations).*

category, words starting with a certain sound, or words rhyming with a target word) within a time limit of 30 seconds. Semantic fluency in particular is very close to what is measured in the productive elements of the German test, and for the purposes of this example, will stand as the definitional measure of Word Finding Difficulties. Figure 5 depicts the performance levels and variability on the fluency task for a sample of 31 children with WFD drawn from one of the first studies to seek an objective definition of this behavioural disorder (Dockrell *et al.*, 2001), and a sample of 12 children and adults with WS (Thomas, Grant, Ansari, Parmigiani, Ewing & Karmiloff-Smith, 2002). These data are compared to the expected level given the mean chronological ages of the individuals involved (solid horizontal lines). Given the post-hoc nature of the comparison between the two

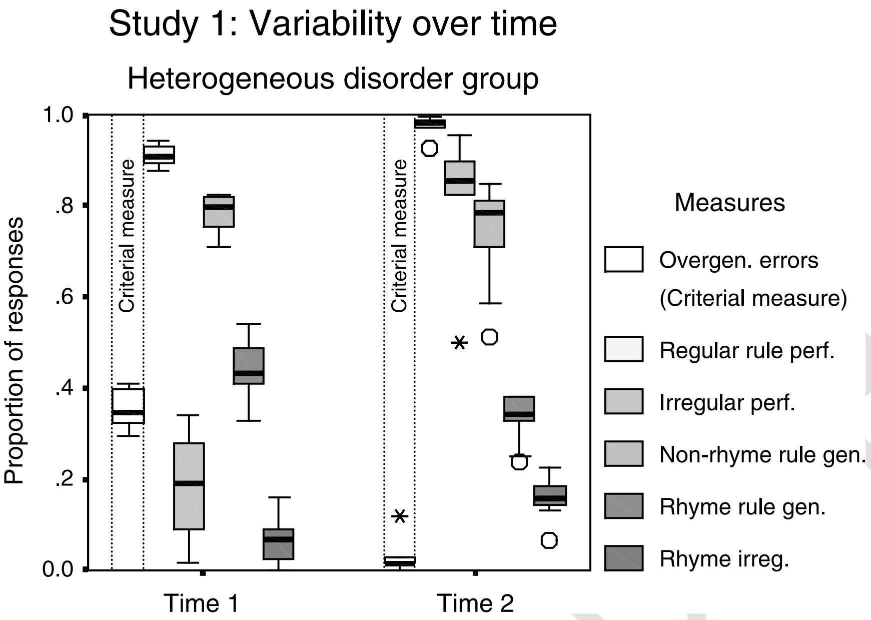

Figure 3 *Continued.*

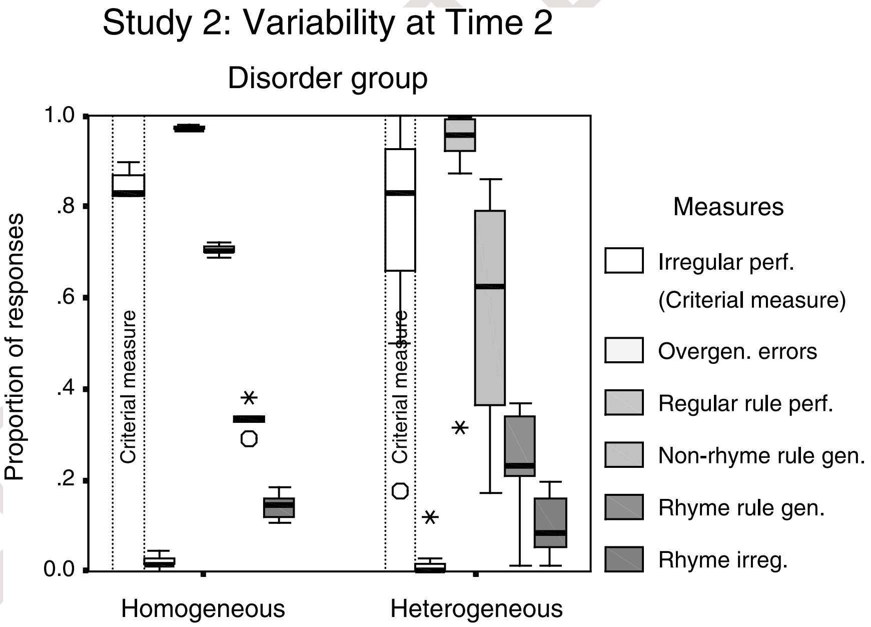

Figure 4 *Boxplots showing the variability of the measures for the homogeneous and heterogeneous disorder groups at the end of training in Study 2. (Compare with Figure 2 to see change over time.)*

disorder groups, it was unfortunately not possible to balance their ages and ability levels. Nevertheless, Figure 5 shows that both groups exhibit a fluency deficit compared to CA controls. Moreover, the variability *is smaller in the WFD group than the WS group*. In each of the three fluency categories, the variance is numerically smaller for the WFD group than the WS group. However, within each group, variability is roughly similar across the different fluency categories, even though different numbers of words are produced for each category.

Table 2 *Within-measure comparisons of variability over time*

	Measure: Group		1	2	3	4	5	6
Study 1	Homogeneous	t1	.062	.010	.021	.011	.045	.021
		t2	.017*	.005	.026	.010	.026*	.035
	Heterogeneous	t1	.045	.022	.115	.042	.061	.056
		t2	.045	.026	.166	.110*	.053	.041
Study 2	Homogeneous	t1	.021	.060	.010	.012	.032	.018
		t2	.028	.014*	.004	.010	.022	.025
	Heterogeneous	t1	.060	.107	.252	.307	.188	.039
		t2	.259*	.037*	.206	.238	.101*	.063
Control		t1	.012	.005	.031	.013	.051	.043
		t2	.000*	.000*	.000*	.029	.046	.028

Note. Scores show standard deviations (* = difference between measure at t1 and t2 significant at 0.05 level, using the Levene test for homogeneity of variances (Levene, 1960)).

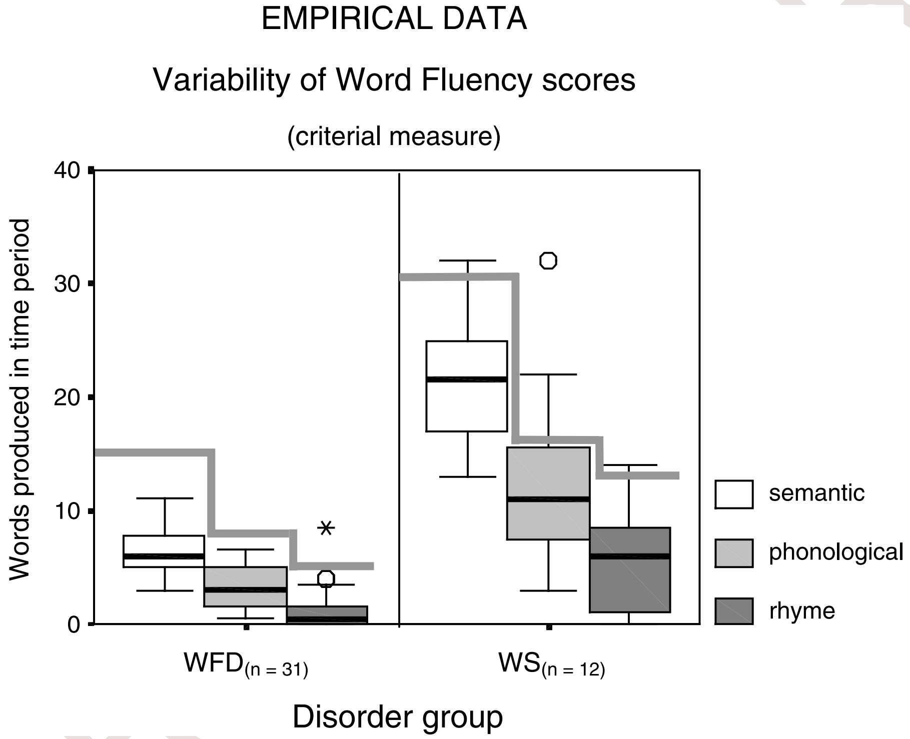

Figure 5 *Empirical data on word fluency for two developmental disorders. WFD = children with Word Finding Difficulty (Dockrell* et al.*, 2001). WS = children and adults with Williams syndrome (Thomas, Grant* et al.*, 2002). The fluency test requires the individual to produce as many words as possible within a time limit, according to either a semantic criterion, a phonological criterion, or a rhyme criterion. For the purposes of this example, fluency is taken as the criterial measure defining the Word Finding Difficulty group. The solid line marks expected performance for chronological-age-matched control groups.*

This implies that the variability is not a direct consequence of the different absolute number of words produced by each group.

The next step is to compare the variability on these 'definitional' scores to additional behavioural measures. The additional measures are taken from another task used to assess the cognitive processes within the language systems involved in productive vocabulary. They focus on speeded naming abilities, and in particular, on the effect of different semantic categories (objects versus actions) and of frequency on the accuracy levels and latencies achieved in picture naming (see Dockrell *et al.*, 2001; Thomas, Dockrell *et al.*, 2002). Figure 6(a) and 6(b) indicate, respectively, the naming accuracy and latency levels for the two disorder groups. Separate comparisons to CA-matched controls indicate that both groups were slower and less accurate than would be expected for their age. Of particular interest, here, is that

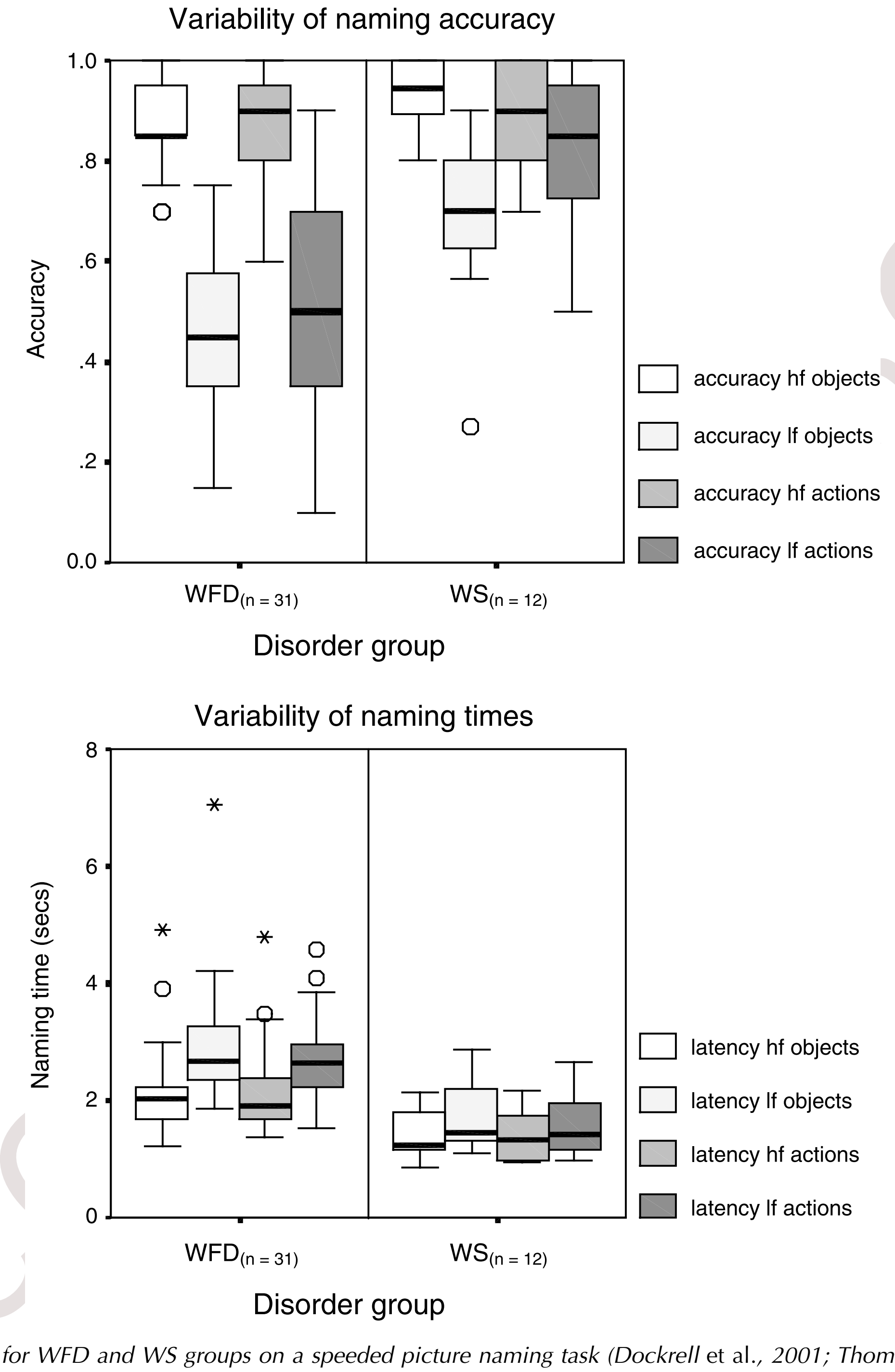

Figure 6 *Variability for WFD and WS groups on a speeded picture naming task (Dockrell* et al.*, 2001; Thomas, Dockrell* et al.*, 2002). (a) Accuracy levels; (b) Naming times. HF = high frequency words, LF = low frequency words. The WFD group shows larger variability than WS group on these measures but smaller variability on the criterial measure.*

both Figures 6(a) and 6(b) now suggest *greater levels of variability in the WFD group than the WS group*. Indeed the behaviourally defined WFD group shows numerically larger variances than the genetically defined WS group in seven of the eight measures, the exception being the accuracy of naming low frequency objects where variances are equal. While variability levels differ across the additional measures within each group, in part due to ceiling/floor effects (e.g. high frequency words show lower variability), *cross-group comparisons* are suggestive of heterogeneous underlying causality in the WFD group.

The comparison used in this example is not ideal because of the different levels of performance of the two groups. It is possible that the lower level of performance of the WFD group contributes to the differential variability. Prospective studies are obviously in a better position to balance disorder groups on performance levels. Nevertheless, this example offers an illustration of how one might apply the heuristic of cross-measure variability to the study of behaviourally defined developmental disorders: if each new (theoretically domain-relevant) measure of the target domain beyond the definitional measure increases the variability across the disorder group, this is not a good sign regarding its causal homogeneity.[3]

General discussion

A computational model of development has been used to derive heuristics that might distinguish between disorder groups with a homogeneous underlying cognitive cause and developmental disorder groups with heterogeneous underlying causes. The model builds in assumptions concerning the different computational sources of variability that produce individual differences and atypical development.

Three caveats need to be added to clarify the status of these heuristics. First, even assuming that the computational model is a valid one, the relative changes to the variability of scores, across different measures and over time, that distinguished homogeneous and heterogeneous groups were tendencies rather than absolutes. Not all additional performance metrics for the heterogeneous group showed larger variability (in Study 1 the figure was 60%, in Study 2 80%). For groups of individual networks sharing the same computational constraints, not all performance metrics showed reduced variability over time (for the homogeneous disorder group in Studies 1 and 2 the figure was 67%, and for the control group it was 83%). The effects appear as tendencies because the variability in each metric is not solely determined by the source of individual differences and the source of atypical development. The structure of the problem domain itself partially determines the variability across measures and its change over time. More generally, one might expect that in problem domains where individuals receive widely different experience, and in problem domains where many partially adequate solutions are available, the contribution of individual and atypical differences to variability may be reduced, and thus the heuristics are less effective. Conversely, in more uniformly experienced domains where success and failure are unambiguous, the heuristics may be highly effective.

It is also worth pointing out that variability in experimental data may also stem from measurement error. To properly compare variability across groups, for instance, it is important to establish as far as possible that measurement error is equivalent. This requires checking that task demands have an equivalent influence across groups. Given that comparison groups can differ in overall ability levels, it may sometimes be necessary to match disorder groups not just against control groups on MA or CA but on absolute level of performance, as a way of controlling for effective task difficulty. However, the issue of measurement error is likely to be of less significance in comparing variability across different measures *within* disorder groups.

The second caveat is that the simulations depicted the ideal situation of an entirely homogeneous disorder group compared to largely heterogeneous disorder groups. Of course in reality, disorder groups might be homogeneous to different extents. A group might have a majority of individuals with one underlying cause and only a few others with different causes. Here, one would hope that across additional measures and over developmental time, the homogeneous portion of the group would hang together more tightly, and allow the others to be identified as outliers.

The third caveat is that the issue of multiple causality has been explored in the much studied and reasonably well understood realm of past tense formation. The findings on multiple causality using this model are likely to generalize to other domains because they essentially derive from *only three assumptions*: (1) that individual and atypical differences have different sources; (2) that each set of atypical initial computational

[3] In response to this example, one might ask the following – why couldn't the speeded naming results form the definitional measure for WFD? Surely then, with fluency as the 'additional' measure, variability would reduce in the behaviourally defined group from the definitional measure to the additional measure? The crucial point here, however, is that *if speeded naming were the definitional criterion, the WFD group would not contain the same individuals*. Those at the higher end of speeded naming performance would not have been recruited to the group in the first place. And new individuals who may have had poorer speeded picture naming performance but perhaps higher fluency levels would have been added. Thus if the WFD group were indeed causally heterogeneous, defining it according to speeded naming would cause the variability of this measure to go down, and the variability of the fluency measure to go up. On the other hand, since the causally homogeneous WS group was defined independently, its variability would not change.

constraints tends to cause a consistent *pattern* of changes across different behavioural measures within a domain; and (3) that through development, performance tends towards a ceiling in large part determined by system constraints. So long as these assumptions hold, the heuristics may well prove useful across a range of domains.

However, generalizing the finding does raise one significant problem. What constitutes a 'cognitive domain', and what constitute 'additional behavioural measures' within that domain? For some cases, the answer is relatively straightforward: for example, for reading, measures would include reading regular words, reading irregular words, reading words with different frequencies, reading words with different semantic properties, pronouncing nonwords of different sorts, and so forth. For other cases, the answer may be less easily forthcoming. For instance, if a developmental disorder is defined on the basis of poor face recognition, should one include object recognition as an additional behavioural measure of the 'domain'? If a developmental disorder is defined according to poor arithmetic, should one include reading skills or grammar skills as relevant additional measures? This issue pertains not only to the identification of separate cognitive domains, but also to how those domains 'hang together' developmentally. Broadly, however, what are conceived of here for the notion of 'additional measures' are aspects of behaviour that might be expected to be dealt with by the same cognitive system according to current theories; or alternatively, that are dealt with by a system that might be expected to share the same computational constraints as that which processes the target domain (as perhaps in the case of face and object recognition).

In conclusion, however, it is proposed that the simple heuristics identified in this article are an important first check when one studies a behaviourally defined developmental disorder:

- If most additional behavioural metrics applied to the disorder group appear to be generating larger variability than the definitional measure, then there is a good chance that the disorder does not have a single underlying cognitive cause.
- If variability in the disorder group does not appear to be decreasing over time in line with a control group matched on initial levels of performance, then once more there is a good chance that the disorder group does not share a single underlying cognitive cause.

Finally, the great value of the modelling approach is that it permits us to focus on principles without the difficulties associated with uncertain phenotypes and vague behavioural classifications. The fact that conclusions about the character of multiple causality in developmental disorders can be derived from computational modelling supports the idea that such models are likely to play an essential role in helping us to understand how computational constraints affect the success (and failure) of developmental processes.

Appendix A

Simulation details

The Plunkett and Marchman (1991, 1993) model was taken as the baseline system. The model employed an artificial language representative of the past tense domain, comprising 500 triphonemic verb stems created by combining English consonants and vowels into three possible templates conforming to the phonotactics of English. In this model, a connectionist network had to learn to associate the stem of each verb with its past tense form. To date, the Plunkett and Marchman past tense model is the one most thoroughly applied to and evaluated within the developmental framework (see e.g. Marcus, 1995; Plunkett & Marchman, 1996). More recent models have added greater complexity in the form of additional semantic inputs and the requirement to learn multiple inflections (see Thomas & Karmiloff-Smith, 2002b, for a review). Here, the simpler model is retained for simplicity and clarity – although the effect of the manipulations on developmental performance for this simple model was qualitatively the same as in a more complex model that incorporated semantic representations.

Baseline 'normal' model

Training set

The training set was based on the 'phone' vocabulary from Plunkett and Marchman (1991, p. 70). Triphonemic stems mapped either to regular past tenses (410) or one of three types of irregular past tense, arbitrary (2), no change (20) or vowel change (68). For clarity, performance is reported only on the vowel change irregular verbs, although the results were similar for the other two types. Each stem was assigned to be high or low frequency. Frequency was implemented by modulating the level of weight change. High frequency stems were given a frequency of 0.3, low frequency verbs 0.1, with the exception of arbitrary irregular past tenses, which had a high frequency value of 0.9 and low frequency of 0.3. A set of 500 novel verb stems was created to evaluate generalization. Novel stems either shared two phonemes with existing regular verbs or irregular verbs.

Network details

Representations

The input comprised three phonemes and the output comprised three phonemes and an optional inflection. Each of 32 possible phonemes was represented over a vector of 30 binary values. This was based on a 6-bit distributed code based on articulatory features, to which a five-fold noisy copying process was applied to create a more redundant version of the phonological code (see Thomas & Karmiloff-Smith, 2002a, for further details). There were 90 input units and 100 output units in the baseline model.

Architecture

A three-layer feedforward network with 50 hidden units was used.

Training and testing regime

The networks were initialized with connection weights randomized between ±0.5, and then trained by exposure to the entire training corpus for 5000 presentations with a learning rate of 0.01 and a momentum of 0. Pattern presentation was in random order without replacement. Weight changes were calculated using the backpropagation algorithm (Rumelhart, Hinton & Williams, 1986) and the cross-entropy error measure (see Hinton, 1989). Network performance on both the training and generalization sets was tested at 10, 25, 50, 100, 250, 500, 1000, 2000 and 5000 epochs. Disordered networks were 'diagnosed' at 500 epochs. Networks trained for 100 epochs were used as a 'mental-age match' for the disordered networks.

Manipulations

The following changes were made to the initial constraints of the network to produce various forms of disordered development.

Localist representations

Each phoneme in the verb stem/past tense form was represented by a single unit instead of a distributed pattern over a set of units. This served to eliminate similarity between phonemes. This network had 96 inputs and 100 outputs.

Compressed representations

Each phoneme was represented by the original 6-bit distributed code prior to the addition of redundancy. This served to increase similarity between phonemes. This network had 18 inputs and 20 outputs.

Slow learning rate

The network was trained ten times more slowly, with a learning rate of 0.001 rather than 0.01.

Use of mean-squared error measure rather than cross-entropy

The error between the output and target was calculated by a Euclidean distance measure rather than an entropy measure. When output units were very inaccurate (e.g. outputting 1 instead of 0 or 0 instead of 1), this reduced the effectiveness of the learning algorithm in adapting the network weights.

Greater initial weight variance

Initial network weights were randomized within the range ±1.5, three times the normal value. This caused an effective delay in development as large weights that were initially inappropriately set for the required mappings had to be corrected.

Two-layer network

The input and output layers were directly connected reducing the computational power of the network.

Reduced hidden units

The hidden layer contained 10 units instead of 50, creating a bottleneck in the internal representations and reducing computational power.

Reduced discriminability

The sigmoid activation function in the processing units was assigned a temperature value of 4 instead of 1 (see Hinton & Sejnowski, 1986). This reduced the ability of the units in the hidden and output layers to discriminate effectively between small differences in the activation levels they were receiving.

Acknowledgements

Thanks to Annette Karmiloff-Smith for helpful discussions contributing to this paper. Thanks to Julie Dockrell and David Messer for kindly making their raw data available from children with WFD. Thanks also to Julia Grant, Dorothy Bishop and three anonymous reviewers for their many helpful suggestions. This work was supported by MRC Project Grant no. G9809880 and MRC Programme Grant no. G9715642 to Annette Karmiloff-Smith.

References

Badali, S., Izvorski, R., Ozawa, K., Diamond, A., & Ullman, M.T. (1999). Phenylketonuria as a model for investigating the role of dorsolateral prefrontal cortex in language. Paper presented at the 6th Annual Meeting of the Cognitive Neuroscience Society, Washington, DC, April.

Bishop, D.V.M., North, T., & Donlan, C. (1995). Genetic basis of specific language impairment. *Developmental Medicine & Child Neurology*, **37**, 56–71.

Bromberg, H., Ullman, M., Coppola, M., Marcus, G., Kelley, K., & Levine, K. (1994). A dissociation of lexical memory and grammar in Williams syndrome: evidence from inflectional morphology. Paper presented at the 6th International Professional Conference of the Williams Syndrome Association, San Diego, CA.

Bullinaria, J.A. (1997). Modelling reading, spelling, and past tense learning with artificial neural networks. *Brain and Language*, **59**, 236–266.

Clahsen, H., & Almazan, M. (1998). Syntax and morphology in Williams syndrome. *Cognition*, **68**, 167–198.

Cohen, I.L. (1994). An artificial neural network analogue of learning in autism. *Biological Psychiatry*, **36**, 5–20.

Cohen, I.L. (1998). Neural network analysis of learning in autism. In D. Stein & J. Ludick (Eds.), *Neural networks and psychopathology* (pp. 274–315). Cambridge: Cambridge University Press.

Dockrell, J.E., Messer, D., & George, R. (2001). Patterns of naming objects and actions in children with word finding difficulties. *Language and Cognitive Processes*, **16**, 261–286.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Fletcher, J.M., Foorman, B.R., Shaywitz, S.E., & Shaywitz, B.A. (1999). Conceptual and methodological issues in dyslexia research: a lesson for developmental disorders. In H. Tager-Flusberg (Ed.), *Neurodevelopmental disorders* (pp. 271–305). Cambridge, MA: MIT Press.

Fredrickson, N., Frith, U., & Reason, R. (1997). *Phonological Assessment Battery*. Windsor: NFER-Nelson.

German, D.J. (1989). *Test of Word Finding TWF*. Chicago, IL: Riverside Publishing Company.

Harm, M., & Seidenberg, M.S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, **106**, 491–528.

Hinton, G.E. (1989). Connectionist learning procedures. *Artificial Intelligence*, **40**, 185–234.

Hinton, G.E., & Sejnowski, T.J. (1986). Learning and relearning in Boltzmann machines. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Hoeffner, J.H., & McClelland, J.L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E.V. Clark (Ed.), *Proceedings of the 25th Child Language Research Forum*. Stanford: Stanford University Press.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, **2**, 389–398.

Levene, H. (1960). Robust test for the equality of variance. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–295). Palo Alto, CA: Stanford University Press.

Marchman, V.A. (1993). Constraints on plasticity in a connectionist model of the English past tense. *Journal of Cognitive Neuroscience*, **5**, 215–234.

Marcus, G.F. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, **56**, 271–279.

Montgomery, J.W., & Leonard, L.B. (1998). Real-time inflectional processing by children with specific language impairment: effects of phonetic substance. *Journal of Speech Language and Hearing Research*, **41**, 1432–1443.

Moore, M., & Johnston, J. (1993). Expressions of past time by normal and language-impaired children. *Journal of Communication Disorders*, **28**, 57–72.

Newfield, M.U., & Schlanger, B.B. (1968). The acquisition of English morphology by normal and educable mentally retarded children. *Journal of Speech and Hearing Research*, **11**, 693–706.

Oetting, J., & Horohov, J. (1997). Past tense marking by children with and without specific language impairment. *Journal of Speech and Hearing Research*, **40**, 62–74.

Oliver, A., Johnson, M.H., Karmiloff-Smith, A., & Pennington, B. (2000). Deviations in the emergence of representations: a neuroconstructivist framework for analysing developmental disorders. *Developmental Science*, **3**, 1–23.

Pennington, B.F., & Smith, S.D. (1997). Genetic analysis of dyslexia and other complex behavioral phenotypes. *Current Opinion in Pediatrics*, **9**, 636–641.

Pezzini, G., Vicari, S., Volterra, V., Milani, L., & Ossella, M.T. (1999). Children with Williams syndrome: is there a single neuropsychological profile? *Developmental Neuropsychology*, **15**, 141–155.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, **103**, 56–115.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition*, **38**, 1–60.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition*, **48**, 21–69.

Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the English past tense. *Cognition*, **61**, 299–308.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D.E. Rumelhart, J.L. McClelland & the PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, **96**, 452–477.

Sieratzki, J.S., & Woll, B. (1998). Toddling into language: precocious language development in motor-impaired children with spinal muscular atrophy. In A. Greenhill, M. Hughes, H. Littlefield & H. Walsh (Eds.), *Proceedings of the 22nd Annual Boston University Conference on Language Development, Volume 2* (pp. 684–694). Somerville, MA: Cascadilla Press.

Simonoff, E., Bolton, P., & Rutter, M. (1998). Genetic perspectives on mental retardation. In J.A. Burack, R.M. Hodapp & E. Zigler (Eds.), *Handbook of mental retardation and development* (pp. 41–79). Cambridge: Cambridge University Press.

Temple, C., Almazan, M., & Sherwood, S. (in press). Lexical skills in Williams syndrome: a cognitive neuropsychological analysis. *Journal of Neurolinguistics*.

Thomas, M.S.C., Dockrell, J.E., Messer, D., Parmigiani, C., Ansari, D., & Karmiloff-Smith, A. (2002). Naming in Williams syndrome. Manuscript submitted for publication.

Thomas, M.S.C., Grant, J., Ansari, D., Parmigiani, C., Ewing, S., & Karmiloff-Smith, A. (2002). Fluency scores for a sample of children and adults with Williams syndrome. Unpublished data.

Thomas, M.S.C., Grant, J., Gsödl, M., Laing, E., Barham, Z., Lakusta, L., Tyler, L.K., Grice, S., Paterson, S., & Karmiloff-Smith, A. (2001). Past tense formation in Williams syndrome. *Language and Cognitive Processes*, **16**, 143–176.

Thomas, M.S.C., & Karmiloff-Smith, A. (2002a). Modelling typical and atypical cognitive development. In U. Goswami (Ed.), *Handbook of childhood development* (pp. 575–599). Oxford: Blackwell.

Thomas, M.S.C., & Karmiloff-Smith, A. (2002b). Modelling language acquisition in atypical phenotypes. Manuscript submitted for publication.

Thomas, M.S.C., & Karmiloff-Smith, A. (in press, a). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioural and Brain Sciences*.

Thomas, M.S.C., & Karmiloff-Smith, A. (in press, b). Connectionist models of development, developmental disorders and individual differences. In R.J. Sternberg, J. Lautrey & T. Lubart (Eds.), *Models of intelligence for the next millennium*. American Psychological Association.

Tomblin, J.B., & Pandich, J. (1999). Lessons from children with specific language impairment. *Trends in Cognitive Sciences*, **3** (8), 283–285.

Ullman, M.T., & Gopnik, M. (1999). Inflectional morphology in a family with inherited specific language impairment. *Applied Psycholinguistics*, **20**, 51–117.

Van der Lely, H.K.J. (1999). Learning from grammatical SLI. *Trends in Cognitive Sciences*, **3** (8), 286–287.

Van der Lely, H.K.J., & Ullman, M.T. (2001). Past tense morphology in specifically language impaired and normally developing children. *Language and Cognitive Processes*, **16**, 177–217.