Using developmental trajectories to understand genetic disorders

Michael S. C. Thomas[1], Dagmara Annaz[2], Daniel Ansari[3], Gaia Scerif[4], Chris Jarrold[5], and Annette Karmiloff-Smith[1]

[1]Developmental Neurocognition Laboratory, School of Psychology, Birkbeck College, London, UK

[2]Department of Human Communication Sciences, University College London, UK

[3]Department of Psychology, University of Western Ontario, Canada

[4]Department of Experimental Psychology, University of Oxford, UK

[5]Department of Experimental Psychology, University of Bristol, UK

Running head: Developmental trajectories and disorders

Address for correspondence:

Dr. Michael Thomas
Developmental Neurocognition Laboratory
School of Psychology
Birkbeck College, University of London
Malet Street, Bloomsbury
London WC1E 7HX, UK
Email: m.thomas@bbk.ac.uk
Web: http://www.psyc.bbk.ac.uk/research/DNL/
Tel.: +44 (0)20 7631 6386
Fax: +44 (0)20 7631 6312

Abstract

We compare two methodologies for studying language and cognitive impairments in developmental disorders: *developmental trajectories* and *matching*. We assess the theoretical frameworks with which they are often associated, as well as their strengths, limitations and practical implications. The contrast between the methodologies is highlighted using the example of *developmental delay* and the criteria used to distinguish delay from atypical development (sometimes called deviance). We argue for the utility of the trajectory approach, using illustrations from studies investigating language and cognitive impairments in individuals with Williams syndrome, Down syndrome and Fragile X, as well and high-functioning and low-functioning children with autism. We conclude that (a) an understanding of mechanism will be furthered by the richer descriptive vocabulary provided by the trajectories approach (for example, distinguishing different types of delay that are conflated in the matching approach); and (b) an optimal design for studying developmental disorders is to combine initial cross-sectional designs with longitudinal follow-up.

When researchers investigate behavioural deficits in individuals with developmental disorders, a common methodology is to proceed as follows. The disorder group is matched with two separate typically developing control groups, one based on chronological age (CA) and a second based on mental age (MA) derived from a relevant standardised test. If the disorder group shows an impairment compared to the CA-matched group but *not* the MA-matched group, individuals with the disorder are taken to exhibit developmental delay on this ability. If, by contrast, the disorder group shows an impairment compared to *both* control groups, then they are taken to exhibit developmental deviance or atypicality (see, e.g., Hodapp, Burack & Zigler, 1990).

Recently, an alternative methodology has been increasingly applied to the study of disorders based on the idea of *developmental trajectories* or *growth models* (Annaz, Karmiloff-Smith, & Thomas, in press; Ansari, Donlan & Karmiloff-Smith, in press; Jarrold & Brock, 2004; Karmiloff-Smith, 1998; Karmiloff-Smith et al., 2004; Rice, 2004; Rice, Warren & Betz, 2005; Scerif, Karmiloff-Smith, Ansari, & Tyler, submitted; Singer Harris, Bellugi, & Bates, 1997; Thomas et al., 2001, 2006). In this alternative approach, the aim is to construct a function linking performance with age on a specific experimental task and then to assess whether this function differs between the typically developing group and the disorder group.

Does it make a difference which methodology is used to study developmental disorders? Does data collection fashion the resulting theory? In this article, we review and compare the *matching* and *developmental trajectory* methods. To anchor our discussion, we contrast the two methodologies in the context of the notion of *developmental delay*.

The concept of delay is widely used in the study of developmental disorders as a method to classify children's cognitive abilities, but in some ways the concept is a

problematic one. Elsewhere, we argue that the notion of delay runs the risk of being descriptively inadequate and explanatorily empty (Thomas et al., 2007). For example, although delay is often used as if it were a mechanistic explanation, it sometimes amounts to little more than a re-description of behavioural data that indicates that the disorder group has produced similar scores and errors to younger typically developing controls. There is no additional elaboration of the causal mechanisms by which this similarity may have arisen. If delay were a causal mechanism, one might imagine that some straightforward predictions should follow. If delay only serves to modulate the rate of development in the cognitive system, performance in the disorder group should eventually reach the same endpoint as in the typical population; and on grounds of parsimony, the delay should be the same across all cognitive domains. Yet in many cases, neither pattern is observed in those individuals who are described as having developmental delay (see Thomas et al., 2007, Karmiloff-Smith et al., 2003, for further discussion). For current purposes, a focus on delay provides the opportunity to illustrate how developmental trajectories can be utilised to explore developmental deficits; and in turn, the use of trajectories demonstrates how the label 'delay' in fact encompasses several different behavioural patterns that may ultimately require different mechanistic explanations. Our focus here will be on improving the descriptive adequacy of the idea of developmental delay.

We begin our comparison by reviewing the traditional methodology used in the empirical investigation of disorders such as developmental dyslexia, Specific Language Impairment, autism, Down syndrome, Williams syndrome, Velo-Cardio-Facial syndrome, Turners syndrome and Fragile X syndrome. We then discuss the developmental trajectory approach and show how it can delineate different forms of delay. In two further sections, we illustrate the use of trajectories with a number of

examples drawn from our own studies, and consider practical issues that arise in their use, such as interpreting null findings, dealing with variability, and validating cross-sectional trajectories via longitudinal follow up. We finish by examining how the two methodologies allow us to decide whether or not a given pattern of development can be classified not as delayed but as qualitatively atypical (deviant, disrupted) – a distinction that many have argued is key in the study of developmental impairments of language and cognition.

Methodology 1: Individual or group matching

The use of CA-matched and MA-matched control groups to study developmental deficits has its origin in a theoretical debate on learning disability (or mental retardation, to use the US terminology) that contrasts the *developmental* and *difference* stances (e.g., Bennett-Gates & Zigler, 1998; see Hodapp & Zigler, 1990, for discussion of the debate in the context of Down syndrome). *Difference* theorists view learning disability as caused by underlying organic dysfunction, producing specific deficits in cognitive functioning and qualitatively atypical cognitive development. By contrast, *developmental* theorists view this characterisation as only applying to a subset of individuals; additionally, there will be a group of individuals with learning disability who fall at the extreme lower end of the distribution of normal individual variation. These individuals will show the same overall qualitative pattern of development as non-impaired individuals, including a similar sequence of developmental milestones and a similar structure to their intelligence (Bennett-Gates & Zigler, 1998). Although, by definition, one would expect the disorder group to exhibit impairments compared to CA-matched controls, an extreme-normal-variation

group should look indistinguishable from a group that is individually matched on a mental-age measure that indexes the stage of developmental progression.

The *development* and *difference* positions identify developmental processes in different sorts of individual. However, the dichotomy is often applied to different component cognitive abilities within the same individual. For example, Figure 1 depicts the type of data that is often reported using this method (usually analysed using t-tests, analyses of variance, or chi-squared tests). In the example shown, performance is contrasted on two tasks to assess whether a developmental dissociation is present, perhaps to test a theory that the abilities tapped by the two tasks develop independently. In Figure 1, the disorder group performs at a lower level than the CA-matched group on both tasks. On Task A, the disorder group performs in line with MA-matched controls, while on Task B there is a deficit compared to MA-matched controls. The results would be interpreted as follows: the disorder group is impaired / atypical / deviant on Task B, while on Task A they are delayed[1] rather than impaired. Where the experimental tasks tap areas of weakness in a disorder, individuals with the disorder are expected to perform below the level of CA controls, and so this latter control group is sometimes omitted (see e.g., Clahsen & Almazan, 1998; van der Lely & Ullman, 2001).

==========

Insert Figure 1 about here

==========

There are two ways in which control groups can be matched to the disorder group. One can seek to carry out *individual* matching, where for each individual in the disorder group, a typically developing individual is selected with the same CA or MA;

[1] For some reason, the term is often qualified by 'merely', 'simply', or 'just'. Sometimes, behaviour in line with MA-matched controls is described as 'intact', 'spared', or 'preserved', potentially obscuring the fact that performance is not at CA-appropriate levels (see later).

or one can be content that the mean CA or MA of the entire control group matches the mean CA or MA of the entire disorder group. *Group* matching is less desirable if the distribution of ages or abilities differs between control and disorder groups since spurious differences in behaviour could arise from this disparity, while individual matching inserts a selection requirement that may reduce the generalisability of the findings (Mervis & Robinson, 2003). Group matching is less demanding on recruitment and may be adopted for practical reasons. Hereafter, we will combine these two methods and refer to them jointly as the *matching* approach.

Designs with MA-matched control groups rely on the use of standardised tests to match the level of developmental progression in the disorder group. This necessarily means that the group comparison is theory dependent: it is important for experimenters to be aware that they are taking a theory-driven view on what standardised test adequately measures developmental progression in the domain that the experimental task is thought to tap (from the range of standardised tests available) (see Yule, 1978). [2] For example, in tasks exploring disorders of language development, the experimenter might match the MA-group according to standardised tests of receptive vocabulary, or productive vocabulary, or receptive grammar. In a typical receptive vocabulary test, the individual has to point to one of four pictures that corresponds to the word they have heard. But it is a theoretical assumption that performance on such a standardised test is the correct single measure to assess developmental progress for, say, a task exploring semantic priming in visual word recognition. One alternative is to use composite MA measures that average across a set of standardised tests to produce a 'verbal' MA or even a 'global' MA. However,

[2] There is the additional issue that the matching task and the experimental task may differ in their task demands. Performance differences between the MA-matched control group and the target group could then arise from different responses to the task demands rather than (or as well as) the cognitive process being measured.

frequently the point of investigating a given disorder is that performance is unequal across cognitive domains or even within domains (e.g., within language, between vocabulary and grammar). By contrast, the control group will tend to have more closely correlated abilities on all the subtests. The result of composite MA measures can be a control group that exceeds the ability of the disorder group on some standardised measures but falls short on others, compromising the interpretation of any task differences (Jarrold & Brock, 2004; cf. Klein & Mervis, 1999). The choice to select an MA group according to a composite measure is another theoretically driven decision made by the experimenter.

Once a theory-driven decision has been made about an appropriate MA group and once the data have been collected, there is a sense in which the experimenter is committed to this theoretical position. There is little flexibility to employ alternative measures of MA. One response to this is to recruit multiple MA-matched control groups using different measures of MA, one per theory about which standardised test is relevant, with an attendant increase in the size and costs of the experiment. This approach may generate multiple conclusions about delay and deviance, if some MA-matched groups are equivalent in their performance to the atypical groups while others are in advance or fall behind the experimental groups. This creates another situation in which the experimenter must commit to a particular theory about the result that provides the most meaningful reflection of performance differences and similarities between groups. This multiple MA-group technique is nonetheless common in research on disorders of language and reading development.

In practical terms, the matching method must avoid floor effects or ceiling effects on the task measures and standardised tests, since these render interpretation of results difficult or impossible (Strauss, 2001). For example, if a participant is at floor,

his or her real ability level is unmeasured because we do not know how far below floor the ability level falls – the measure is no longer working. Preferably, the CA, MA, and disorder groups should all be in the sensitive range of the tests and, at the very least, the MA and disorder groups should be in the sensitive range. This may limit the matching technique where individuals with disorders have severe deficits because there may be no age-equivalent performance in the typically developing population. The methodology is optimal when the disorder group covers a very narrow age range, and/or when the experimental measure is only sensitive around a particular age. It is less advantageous when groups are averaged over a wide age range, which can sometimes be the case in studies of rare developmental disorders. This is because group mean performance may mask a fairly wide range of performance, again limiting interpretability and inference to causal mechanism.

MA matching relies on the use of age-equivalent scores from standardised tests. For a given test score, one derives the age at which the average child from the (typically developing) standardising population achieved this score. Several limitations have been noted in these tests (McCauley & Swisher, 1984). For example, age-equivalent scores are silent on the variability present in the standardising population at each age. Many of the typically developing children may have scored some way below (or above) the average age-equivalent score in the standardisation sample, yet disparities of this nature are not treated as deficits (or hyper-functioning) as they are in disorder groups.

Finally, one often ignored but crucial consequence of the matching method is that although it is being used to study (potentially atypical) development – that is, how behaviour and cognition change with age – age is actually factored out at the design stage. Age is not a variable but a label assigned to a control group. As such, it

is treated in the same way that one would treat a confounding variable. It is conceivable that de-emphasising age in this way has a consequence for the types of theories generated from these studies.

Methodology 2: Developmental trajectories

The aim of the developmental trajectory approach is twofold. Firstly, it aims to construct a function linking performance with age for a specific experimental task, and then to compare the respective functions of the disorder group and a typically developing group. Secondly, it aims to establish the developmental relations between different experimental tasks, assessing the extent to which performance on one task predicts performance on another task across development and once more, to compare the developmental relations found in the disorder group with those observed in a typically developing group. In an ideal world, both comparisons would comprise longitudinal group studies. However, the method is also applicable to cross-sectional studies or a combination of the two (see later).

The use of trajectories in the study of developmental disorders has its origin in growth curve modelling (see, e.g., Rice, 2004; Rice, Warren, & Betz, 2005; Singer Harris et al., 1997; Thelen & Smith, 1994; van Geert, 1991) and in the wider consideration of the shape of change in development (see, Elman et al., 1996, chapter 4). The impetus to move from matching to trajectory-based studies was motivated by a concern that explanations of developmental deficits based on the matching approach were becoming increasing non-developmental in nature (see Karmiloff-Smith, 1998, for discussion). Developmental behavioural impairments were frequently being explained with reference to static, non-developmental, and even adult models of cognition (see, for example, Thomas & Karmiloff-Smith's (2002) discussion of

Coltheart et al.'s (1993) explanation of developmental dyslexia). Such theoretical leanings were not a consequence of the matching methodology per se, although they were perhaps encouraged by the exclusion of age as a variable in the design. Instead, the theoretical leanings were driven by an implicit extension of the explanatory framework of adult neuropsychology to the developmental realm. For example, were Figure 1 to depict data from a study of an adult acquired deficit, one might interpret the disorder group's impairment on Task B compared to MA-matched controls in terms of a modular cognitive system in which there had been focal damage to the mechanism responsible for Task B. However, for developmental disorders, this explanation ignores the fact that the behavioural deficit is the outcome of an adaptive, developmental process likely to be characterised by features such as interactivity, compensation, and redundancy (Bishop, 1997; Karmiloff-Smith, 1997, 1998; Thomas, 2007). Moreover, the modular structure identified in normal adulthood is unlikely to be a precursor to development (Paterson et al., 1999).

The trend for non-developmental explanations can be observed by the loose appropriation of terminology from the study of adult brain damage to describe developmental deficits. Cognitive mechanisms are labelled as 'intact', 'spared', or 'preserved' when what is meant is that they are developing normally, and described as 'impaired' or 'damaged' when what is meant is that they are developing atypically. By not couching the explanation of normal behaviour as a proposal in terms of normal developmental process, the terminology effectively overlooks the possibility that normal-looking behaviour might be produced by atypical process in a developmental disorder (Karmiloff-Smith, 1998). By contrast, the use of trajectories makes a more explicit appeal to the researcher to explain his or her data in terms of *change over time*

or *developmental relations* between cognitive processes, and in terms of a (potentially atypically) constrained developmental process.

Constructing trajectories

How might an appeal to trajectories reverse this trend? Let us begin by examining how trajectories are constructed. We first consider functions that link performance with chronological age and the comparisons with typically developing controls that this permits. We then consider developmental relations and functions that link performance with mental age, which may serve as a more stringent test of delay/deviance hypotheses.

For a cross-sectional design, the trajectory method works as follows. A disorder group is recruited in which there is a reasonable developmental age range (i.e., spanning childhood, adolescence, and adulthood, but not adulthood alone). Performance is assessed on the experimental task. Additionally, standardised test results are collected on as many measures as are thought relevant to the cognitive process under study (within limits of practicality). A typically developing comparison sample is then recruited that spans from the youngest mental age of the disorder group on any of the standardised measures to the oldest chronological age, and the performance of these comparison individuals is assessed on the experimental task. The approach relies on using an experimental task that will yield sensitivity across the ability range of the disorder group, avoiding floor and ceiling effects where possible.

The analysis begins by constructing a task-specific developmental trajectory for the control group, using regression methods to derive a function linking task

performance with age.[3] We will mostly assume the use of linear methods, since these aid in understanding the relationships between trajectories (see next section). This may mean transforming either age or the dependent variable or both to improve linearity. Figure 2 shows an illustrative set of results for a typically developing group and a disorder group. The figure depicts all the individual data, reflecting one of our preferences in using the trajectory approach (see later Table 2).

There are now three types of comparison that can be made between the disorder group and the typically developing (TD) trajectory. The first type of comparison is *theory neutral*. Here, the researcher merely asks whether the performance of each individual in the disorder group can be fit anywhere on the TD trajectory. If the experimental task only has a single dependent variable, this may not be a particularly useful comparison. That is, if TD performance stretches from 0 to 100% on some measure, it is evident that any individual can be fit on that trajectory. The comparison is in fact tantamount to standardising your own experimental task, so that a mental age measure can be derived for each individual in the disorder group (the mean age of the TD sample at which a given performance level is exhibited). However, when the experimental design includes two or more measures (e.g., performance on high frequency versus low frequency items), the theory neutral comparison can be much more informative. The researcher can ask whether a given disparity between the two measures (e.g., the frequency effect) for an individual with the disorder is observed *anywhere* on the TD trajectory. If it cannot, here is a theory-neutral marker of atypicality. (Strictly speaking, it is theory neutral in respect of the comparison; there is a theory in the experimental design that the relationship between

[3] Linear regression may be approximated by splitting the age range into several groups and using an analysis of variance with a multi-level age factor (see Ansari, Donlan & Karmiloff-Smith, in press).

the two measures, such as performance on high and low frequency items, should be developmentally robust).

==========

Insert Figure 2 about here

==========

The second type of comparison now allows for the construction of a trajectory for the disorder group, linking their performance on the experimental task with their *chronological age*. This trajectory can then be compared with the TD trajectory to assess whether the disorder group shows a difference in their developmental performance on the task. While this is likely when studying areas of weakness in the disorder, it is a more open question for cognitive domains outside the primary deficit (such as non-verbal abilities in children with Specific Language Impairment). For a single dependent variable, the comparison of two trajectories will involve a linear regression model with one between-groups factor. For multiple dependent variables (such as in the example of the frequency effect), this will involve a mixed-design linear regression model including within-participants factors to compare several trajectories simultaneously.[4] Confidence intervals around the regression line can be used to assess the age at which trajectories converge or diverge. Figure 2(a) depicts data for the CA-based comparison. Note that the TD group extends to a younger age, and in this case, the disorder group appears to have a lower level of performance and to be developing more slowly.

[4] The SPSS Univariate General Linear Model function can be adapted to perform between-group linear regression. Similarly, the SPSS Repeated Measures General Linear Model function can be adapted to perform mixed-design linear regression that includes within-participant factors. Both functions allow evaluation of overall fit of model, influence of outliers, and measures of effect size. (See www.psyc.bbk.ac.uk/research/DNL/Thomas_trajectories.html for sample data and worked examples of trajectory analyses using SPSS).

The third type of comparison considers *developmental relations* in the disorder group. A separate trajectory can be constructed for each standardised test measure collected from the disorder group, in which a function is derived linking the mental age (test age equivalent) on that test with task performance. Each mental-age trajectory can then be compared against the TD trajectory. If task performance is in line with a given standardised measure, then plotting the disorder group's data according to each participant's MA should move the atypical trajectory to lie on top of the TD trajectory.

More sophisticated comparisons are possible. For example, one can use the TD trajectory to standardise the performance of the members of the atypical group. Let us say that the experimental task was some aspect of morphology and one had collected standardised scores for the disorder group on a receptive vocabulary test as a measure of their verbal MA. One can then derive a residual score for each individual in the disorder group based on the difference between their observed task score (e.g., on the morphology task) and the score predicted by their MA, according to the TD trajectory (see Jarrold & Brock, 2004). These residuals can be standardised to create z-scores that can be compared across different experimental tasks. Thus one could derive z-scores for the disorder group on a syntax task and ask whether, on the basis of their verbal MA, are there disparities in the expected levels of morphology and syntax. Comparisons are possible across different experimental tasks (e.g., morphology, syntax) standardised on the same MA measure (e.g., receptive vocabulary) or across the same task (e.g., morphology) under standardisations based on different MA measures (e.g., a receptive vocabulary test and a receptive grammar test) (see, Jarrold, Baddeley, & Phillips, in press, for details of these methods).

As long as there is an opportunity to collect multiple standardised test results on the disorder group, the trajectory method gives great flexibility at the analysis stage to evaluate potential relationships to the TD trajectory. This contrasts with the matching approach, where a decision is made at the design stage to recruit an MA-matched control group based on a particular standardised test. Usually, a larger number of TD controls will be collected in the trajectory approach with a weaker selection bias, giving a fuller picture of typical development on the task. Figure 2(b) depicts performance plotted against an MA measure. For these illustrative data, it becomes evident that the disorder group has a lower level of performance than the TD group even when their lower MA is taken into account but now the disorder group is developing at the same rate. Results of this type would suggest that, to the extent that the standardised test is a valid index of development in the target cognitive domain, the delay is uneven across component processes.

Note that the use of simple correlations to explore developmental relations between cognitive abilities in disorders effectively falls within the trajectory approach. However, when researchers use simple correlations, they do not always plot these trajectories to illustrate the degree of variability, or establish the linearity of relationships between abilities, or check the influence of outliers on the relationship, or the presence or absence of ceiling and floor effects, and so forth. In our view, the more explicit use of trajectories is therefore preferable when relationships are explored.

The trajectory method is advantageous where there is a wide age (and potentially ability) range in the disorder group. For the study of developmental relations within the disorder group, a wide ability range is the important consideration. The trajectory method relies on using test measures that have

sensitivity across the wide age range. It may therefore appeal to dependent variables such as reaction time rather than just accuracy, and use implicit rather than explicit measures of performance (Karmiloff-Smith et al., 1998). These features contrast with the matching approach, which is ideal for narrow age ranges and can tolerate a test with a narrow sensitive range, as long as that range is appropriate for the ability of the disorder and control groups sampled. In common with the matching approach, floor and ceiling effects should be avoided, particularly in the disorder group (see later examples for problems that can arise if floor and ceiling effects are present). And where standardised tests are used to derive mental ages, similar caveats apply regarding the way age-equivalent scores mask potential variability in the TD group (McCauley & Swisher, 1984). The similarities and differences between matching and developmental trajectories methodologies are summarised in Table 1.

==========

Insert Table 1 about here

==========

How does the choice of methodology affect theory, then? As we have indicated, there is no necessary influence. However, the trajectories approach foregrounds behavioural change over time while the matching approach allows group differences simply to be characterised as an impairment. To the extent that methodologies interact with theories, behavioural change focuses the spotlight on the role of the developmental process in producing group differences.

Using trajectories to distinguish types of developmental delay

We are now in a position to consider how trajectories may be useful for studying developmental delay. Under the matching approach, a cognitive ability in a disorder

group is described as delayed if performance falls below the CA-matched control group but resembles that of a control group matched on a mental age deemed relevant for the target cognitive domain. The thrust of this section is that, when construed in terms of developmental trajectories, the performance of the disorder group can resemble that of the younger TD group *in more than one way*. We believe that one of the reasons neurocognitive explanations of delay are thin on the ground is that delay is not sufficiently detailed as a descriptive term. In this section, we show how the use of trajectories distinguishes at least three forms of delay, and how additional descriptors may also discriminate patterns of development that index different underlying causal mechanisms.

Since our terminology will make reference to linear regression equations, we begin by briefly recapping some basics of this method. The use of linear methods assumes that a putative relationship between age (or mental age) and task performance is either linear or can be made to resemble a linear function by transforming the age variable, the dependent variable, or both. By linear, we mean that task performance is a weighted combination of age plus some constant. Linear relationships can be captured by the equation

$$y = ax + b$$

For a trajectory, the equation becomes

$$Test\ Performance = a \times \left(Age\ in\ months\right) + b$$

where *a* and *b* are constants corresponding to gradient (how quickly performance improves) and intercept (the level at which it started), respectively. In a linear system, a change in input (δx) leads to an *identical* change in output ($\delta y = a \times \delta x$) no matter where it occurs in the range of input values (values of x). For the trajectory, an age difference should correspond to the same performance difference at all points across

the age range. Trajectories for which this is not true are called *non-linear*. The use of linear methods is a simplification, but one that makes interpretation of interaction terms more straightforward in more complex designs. However, alternative non-linear regression methods may also be used and there is a diverse range of such functions available to characterise change over time (see Elman et al., 1996).

Linear regression methods derive the function linking two variables from pairs of values (e.g., age, performance) and under sampling assumptions, confidence intervals can be generated around the line indicating the region within which the trajectory is likely to fall with a given level of confidence. In line with standard regression techniques, the first step is to ensure that it is appropriate to fit a trajectory to a data set, so that (i) the trajectory captures a significant amount of the variance, assessed by the R^2 value (e.g., Figure 3(a) and (b) depict two linear regression fits, only one of which corresponds to a reliable trajectory; R^2=.714 for the trajectory in 3(a) [$F(1,97)=240.0$, $p<0.001$], R^2=.016 for the trajectory in 3(b) [$F(1,97)=1.6$, $p=.214$]); (ii) the relationship observed in the data is roughly linear; and (iii) no outlier exerts undue influence on the trajectory (outlier influence can be assessed using measures such as Cook's distance; Cook; 1977).

==========

Insert Figure 3 about here

==========

One difficulty with the linear regression analysis is that a non-significant relationship may arise under two conditions: (1) when the distribution of performance scores is random with respect to the predictor of age, and (2) when the points are distributed horizontally (see Birdsong, 2005, for a similar point). In both cases, age is not useful in determining performance. However, in the second case, provided the

task measure is in the sensitive range and it has been established that the typically developing group improves across the age range, it is possible that individuals with the disorder have indeed progressed as far as they can, and the trajectory has a gradient of zero. Figure 3(c) illustrates idealised versions of the two cases, which we refer to as *no systematic relationship* and *zero trajectory*, respectively. In both cases, the best-fit regressions line are flat; for one trajectory, the best fit line lies in the middle of a random data cloud; for the other, the points are tightly clustered around a narrow performance range across development. We have found it useful to distinguish between these two cases in our research by using a rotation method. Figure 3(d) depicts the same data but transformed by a 45˚ anti-clockwise rotation in geometric space. When the analyses are repeated on the rotated data, the *zero trajectory* now produces a highly significant regression (the R^2 value changes from .0011 to .9999 following rotation) while the *no systemic relationship* produces a similar degree of fit before and after rotation (R^2 changes from 0.00030 to 0.00004). A trajectory that switches from a non-significant R^2 to a significant R^2 following rotation is suggestive of a *zero trajectory* rather than *no systematic relationship*. The reason for a lack of change with increasing age in the zero trajectory would remain to be interpreted.

Assuming that we have two reliable linear trajectories, one for the TD group and one for the disorder group, these trajectories can now be statistically compared. The test will indicate whether there is a significant difference in the rate (gradient *a* in the above equation) and/or the onset (intercept *b* in the above equation). Importantly, where there is a difference between the two trajectories, *three* different types of descriptive delay can now identified. These are depicted in Figure 4 with illustrative data. In Figure 4(a), there is a significant difference in the intercept. Here, delay is manifested in a *later onset of development*. In Figure 4(b), there is a difference in the

gradient between the two trajectories. Here, delay takes the form of a *slowed rate of development* in the disorder group. In Figure 4(c), there is a difference in both parameters, implying that development has *both a delayed onset and a slowed rate.*

==========

Insert Figure 4 about here

==========

A focus on trajectories allows further descriptors to be attached beyond delay, which may ultimately index different underlying developmental pathways. For example, as we suggested above, the TD group may exhibit a reliable trajectory but the disorder group may exhibit no reliable change in performance with age. Figure 4(d) and (e) illustrate two further types of difference. In the first, a *linear relationship is observed in the TD trajectory,* but a *non-linear trajectory is observed in the disorder group*. In the second, a linear relationship is observed in the TD trajectory and this is initially tracked in the disorder group but the disorder group then asymptotes at a lower level of performance.

These alternative descriptors are assigned when an alternative function gives a significantly closer fit to (i.e., a better explanation of) the data than the linear equation. The R^2 value for a regression model indexes how well the model fits the data (specifically, the proportion of variance explained) and R^2 values can be derived for different functions fitted to the same data (e.g., in the SPSS Regression Curve Estimation facility). A higher R^2 gives a better fit. It is possible to test whether a function is a *significantly* better fit by discounting for the extra parameters available in the more complex equations[5]. To illustrate, linear and non-linear functions were fit

[5] This is done by deriving an F-ratio from the relative increase in the sum-of-squares and the relative increase in the degrees of freedom (information available in the ANOVA table for each regression fit). For regression fits 1 and 2, the equation is

to the disorder trajectory in Figure 4(d). The linear function produced an R^2 of .900 while the logistic function (an s-shaped curve) produced an R^2 of .990. Since both models have the same number of parameters, the latter is the better model and so the disorder trajectory would be classified as non-linear. Similarly, when linear and non-linear functions were fit to the data in Figure 4(e), the linear function produced an R^2 of .943 while a quadratic function (including a variable of age-squared) produced an R^2 of .998. The quadratic has more parameters, so a statistical comparison is necessary to show it is a better model and this was indeed the case ($F(1,2)=70.1$, $p=.014$). The disorder trajectory would therefore be classified as non-linear and given its shape, as exhibiting a premature asymptote. Finally, since non-linear functions also have intercepts, one can characterise a trajectory as separately showing a delayed onset followed by a non-linear trajectory.

How would the matching approach deal with the different types of delay we have described. The illustrative data in Figure 4 allow us to make this comparison by averaging across groups. Figure 4(f) demonstrates the mean performance of the TD group and the disorder groups with each type of trajectory, collapsed over age as would be the case in a group comparison. *Delayed-onset+slowed-rate* produces the lowest mean score and *premature asymptote* the highest, while *delayed onset*, *slowed rate*, and *non-linear* all produce similar scores. The fact that, from the perspective of the matching approach, some of these groups are indistinguishable suggests that for wide age ranges at least, the use of trajectories provide a descriptively more powerful empirical vocabulary.

$$F = \frac{\left((SS1 - SS2)/SS2\right)}{\left((DF1 - DF2)/DF2\right)}$$

where SS stands for sum-of-squares and DF for degrees of freedom. This ratio has DF1-DF2 degrees of freedom for the numerator and DF2 degrees of freedom for the denominator (see Motulsky & Christopoulos, 2004).

Let us amplify this point. Where the individual matching approach encourages a monolithic descriptive partition between 'delay' and 'deviance', the use of trajectories distinguishes at least *seven ways* that a disorder group can statistically differ from a control group in the functions that link performance and age (mental age): (1) *delayed onset*, (2) *delayed rate*, (3) *delayed-onset+slowed-rate*, (4) *non-linear*, (5) *premature asymptote*, (6) *zero trajectory*, and (7) *no systematic relationship with age*. An accurate characterisation of patterns of change is, of course, a necessary precursor to formulating causal accounts of developmental impairments.

This richer taxonomy of developmental delay, with its focus on developmental change and developmental relations, draws some similar conclusions to the recent work of Rice and her colleagues (see, e.g., Rice, 2004; Rice, Warren, & Betz, 2005). For comparison, Rice (2004) suggests that developmental trajectories should be characterised in terms of their onset timing, their acceleration rate, and points of change in their acceleration, and separate trajectories should be established for the delineated subcomponents of the linguistic system. Rice et al. (2005, p.22) place particular emphasis on the utility of *onset* differences in language development, arguing that delayed onset may be a hallmark characteristic across most of the known clinical forms of language impairments.

Examples of the trajectory approach

In this section, we describe four examples of studies that have used the trajectory method to explore potential differences between one or more developmental disorder groups and a typically developing control group. These examples focus either on language development or on the developmental relations between verbal and non-

verbal development. They serve to illustrate a number of methodological points that arise in using the trajectory approach.

1. Inflectional morphology in Williams syndrome. Early published and unpublished studies of language development in Williams syndrome (WS) suggested that these individuals might have greater problems inflecting irregular nouns and verbs than regular nouns and verbs (Bromberg et al., 1994; Clahsen & Almazan, 1998). This is of theoretical interest because performance on inflecting regular and irregular items is taken to index either the involvement of different mechanisms (rule-based vs. associative learning mechanisms) or the influence of different information sources (phonological vs. lexical-semantic) depending on the theory (see Thomas & Karmiloff-Smith, 2003, for a review). However, these initial studies were compromised by small participant numbers and/or the absence of appropriate statistics. Moreover, the most salient characteristic of language development in WS is that its onset is delayed (see, e.g., Meyer-Lindenberg, Mervis & Berman, 2006), and a characteristic of typical development is that irregular inflections are harder to learn than regulars. Might, then, the apparent problem in irregular inflection stem from a delayed onset in language development rather than a specific deficit to some component of the language system? Figure 5 depicts accuracy levels for a past tense elicitation task, where the participant is either required to produce regular past tenses (e.g., talked) or irregular (e.g., drank). The data are from 18 individuals with WS and 46 typically developing controls (Thomas et al., 2001). Two groups and two verb types produced four trajectories, which were analysed with a mixed-design linear regression model.

==========

Insert Figure 5 about here

==========

There are several points to note here. First, the main finding was that when WS performance was plotted against CA, it was not only lower overall than the TD group but there was also a greater difference between accuracy levels on regular and irregular verbs. However, when WS performance was plotted against mental age based on a test of receptive vocabulary (the British Picture Vocabulary Scale (BPVS); Dunn et al., 1997), the difference between regular and irregular verbs was now equated between the groups (i.e., the interaction of verb type and group became non-significant). The apparent 'deficit' in irregular verb performance resolved into a disparity predicted by the developmental state of the wider language system, which holds both for the normal and atypical case. Note, however, that when performance was plotted against BPVS mental age, the WS group remained slightly less accurate overall, producing reliable differences in trajectory intercepts between the groups. The group difference when WS performance was plotted against MA suggests that in this disorder, verb inflections (a part of productive grammar) fell behind the level predicted by receptive vocabulary. This illustrates the importance of the choice of standardised tests which, as we have argued, is a theory-dependent decision concerning the parts of the cognitive system that are predicted to develop in harness.

Second, the data were partially compromised by a ceiling effect in more able participants (especially in the control group), a problem that has affected many studies of inflectional morphology in disorders (Brock, in press). In an attempt to address this problem, the data were linearised by plotting performance against $\frac{1}{(age)^2}$ where age was calculated in months, but clearly it would have been preferable if the test had been in the sensitive range for all participants. Third, in Figure 5 (left panel) depicting

the relationship between performance and CA, there is not a full-overlap between the groups because the TD group was recruited to span from the WS group's lowest MA to the highest CA. The right panel demonstrates that plotting according to receptive vocabulary now produces full overlap. Fourth, standardised tests usually have a maximum age (in this case, 17 years and 6 months). This presents a difficulty in comparing the disorder group against TD at older chronological ages because, obviously, no individual can produce a test age above the ceiling. If the disorder group never reaches ceiling on the standardised test, the difficulty is to some extent resolved by assigning an MA of the ceiling value to any individual in the TD group whose age falls above the ceiling. Note finally that regular verb performance in the WS group showed no systematic relationship with chronological age (R^2=.0065). This is a point we take up in our next example.

2. Picture naming in Williams syndrome. Early work on language development in Williams syndrome also made another interesting claim. Following up on anecdotal reports of the presence of rare or unusual words in the spontaneous language of individuals with WS, some researchers suggested that this reflected atypical structure in their lexicon, and in particular, an attenuated encoding of word frequency (Rossen et al., 1996). Thomas et al. (2006) explored picture-naming reactions times in a sample of 16 individuals with WS and once more compared these times to those of a typically developing group's trajectory (n=16). Pictures varied according to category (object, action) as well as frequency. Did the WS group show the same sized frequency effect as controls? Before we answer that, let us look at data that were collected to give an indication of baseline naming speed. Figure 6 plots the speed with which individuals named the numerals 1-9. These were highly familiar, over-learned

stimuli that were named with 100% accuracy. In the typically developing population, reaction times tend to decrease with expertise according to a power law (Cohen, Dunbar, & McClelland, 1990), so in this case a log-log transform was used to linearise the data.

Figure 6 demonstrates that the control group showed a reliable reduction in naming time with chronological age. However, the WS group did not: there was no systematic relationship between CA and naming time for numbers in the disorder group (R^2<.0001). What are we to make of this finding? In the following section, we will argue that it does *not* mean that performance does not improve with age in these individuals, because the data are from a cross-sectional trajectory. By contrast, when WS performance was plotted according to their receptive vocabulary, a reliable trajectory now emerged, albeit revealing a naming speed that lagged behind receptive vocabulary expectations, exhibiting delays in both onset and rate. Lastly, was there a reduced frequency effect in WS for naming pictures? Were low frequency words named faster than expected? The answer was no. Contrary to earlier suggestions, when developmental relations were assessed using a mixed-design linear regression model, the frequency effect was in line with MA, as measured by a standardised test of receptive vocabulary.

==========

Insert Figure 6 about here

==========

3. Vocabulary vs. pattern construction in autism: spectrum effects. Cross-syndrome cross-task-domain comparisons can be very informative about the atypical constraints operating in developmental disorders. One can begin by making simple comparisons based on the multiple sub-tests of standardised intelligence tests (although one must

acknowledge that in some respects, these tests have limited sensitivity; see Karmiloff-Smith, 1998; Karmiloff-Smith et al., 1998). Figure 7 depicts data taken from Annaz et al. (2007) in their comparison of Williams syndrome, Down syndrome and autism for children between 5 and 12 years of age. Notably, Annaz et al. (2007; see Annaz, 2006) collected data from low-functioning as well as high-functioning children with autism, in order to explore the influence of the spectrum within this disorder. High-functioning (n=16) and low-functioning (n=17) children were assigned to their groups according to the Childhood Autistic Rating Scale (CARS; Schopler, Reichler & Rochen, 1993). This clinical measure is better suited to assessing variations in the severity of autism than the less sensitive Autism Diagnostic Observation Schedule measure (ADOS; Lord, Rutter, DiLavore & Risi, 1999) (see Saemundsen et al., 2003). Figure 7 plots test ages derived from the BPVS (Dunn et al., 1997) and from the pattern construction subtest of British Abilities Scales II (BAS-II; Elliott et al., 1996) against chronological age for typically developing (n=25), high-functioning (ASD-HF), and low-functioning (ASD-LF) groups with autism. The horizon grey line represents floor performance on each test.

==========

Insert Figure 7 about here

==========

As expected, the TD group received test ages very close to their chronological ages (R^2=0.9626 and .9599 for the two tests, respectively). For receptive vocabulary, the ASD-HF group produced a reliable trajectory that was slightly lower (that is, later in onset) than the TD trajectory, although this difference did not reach significance. By contrast, for the ASD-LF group no reliable trajectory emerged, and indeed most of these children were at or close to floor on the vocabulary test. In one sense, this is not

surprising, since one of the markers of severity in autism is the level of language development. However, one might even question whether these data are themselves valid: perhaps the ASD-LF group was simply unable to complete this task given their ability level? Figure 7 (right panel) allows us to address this question. These data reveal the developmental trajectories on the BAS pattern construction task, in which the children are asked to complete geometric puzzles. Here, both groups with autism produced trajectories overlapping with typical development, and indeed the ASD-LF group produced a tighter trajectory than the ASD-HF group (R^2=.8223 and .3511, respectively). These data reveal stark differences in the profile of children at different points of the autistic spectrum. They also demonstrate that cross-task-domain comparisons can shed light on the validity of the respective trajectories. The normal profile on pattern construction increases confidence that the lack of improvement on vocabulary in the ASD-LF group is a real phenomenon.

4. Verbal and visuo-spatial memory in Williams syndrome and Down syndrome. In this example, we consider some more sophisticated techniques to compare developmental relations between abilities in two disorders. Jarrold et al. (in press) compared the performance of individuals with Down syndrome (n= 20) and with Williams syndrome (n = 15) to that shown by 110 typically developing children on the Doors and People test, a measure of verbal and visuo-spatial recall and recognition memory. Figure 8 plots the performance of these groups on two of the tasks in the battery, the verbal recall and verbal recognition tests. The top two panels of the figure show performance plotted against chronological age, while the lower two panels show performance plotted against verbal mental age (assessed via the BPVS; Dunn et al., 1997). Because of the range of ages and abilities within the typically developing

group, both floor and ceiling effects are apparent, and so the development of performance with age or ability is not linear in this group. Consequently, these regressions were linearised by converting each individual's score into a *probit* score (the z-score corresponding to that individual's score on the task as a proportion of the maximum possible) and then regressing that value against the log of either chronological or verbal mental age. This produced reliable linear fits, and these in turn allowed the authors to determine the extent to which each individual in the atypical groups showed performance that was in line with their chronological or verbal mental age (specifically, the residual scores for each individual were standardised on the basis of these linearised regressions).

==========

Insert Figure 8 about here

==========

Figure 9 shows the resultant standardised residual values under the two different forms of standardisation and indicates how far each disorder group fell below the normal range for recall and recognition. Three key points can be drawn from these data. First, they further emphasise the fact that atypical groups tend to perform poorly on chronological age standardisations, because their abilities lag behind age expected levels. A comparison of the scales of the two graphs in the figure shows that when performance is standardised for VMA, then the atypical groups are much less impaired. Second, when the two atypical groups are standardised for age they perform similarly, yet when compared to TD individuals on the basis of VMA, the individuals with Down syndrome are clearly less impaired than those with Williams syndrome. This reflects the fact that VMA is a relative strength in Williams syndrome and something of a weaker area in Down syndrome; consequently broadly

comparable overall levels of task performance represent different levels of impairment relative to VMA in the two groups. Finally, the figure shows that the type of regression employed to standardise the data has implications for the interpretation of the results. When the groups are standardised relative to chronological age, both perform poorly on both the recall and recognition task. However, under the VMA standardisation the individuals with Williams syndrome show impaired performance on the verbal recall task only. This difference in patterns of impairment reflects the fact that the two tasks are related to chronological and verbal mental age in different ways than in the typical, standardisation sample (see Figure 8).

==========

Insert Figure 9 about here

==========

Other studies using the trajectory methodology can be found in Brock and Jarrold (2004, 2005), Cornish, Scerif and Karmiloff-Smith (in press), Jarrold, Cowan, Hewes and Riby (2004), Jarrold et al. (in press), Karmiloff-Smith et al. (2004), Scerif et al. (2005), Scerif, Tyler, Ansari and Karmiloff-Smith (submitted), and Thomas et al., (submitted).

Practical issues of using trajectories

The above examples indicate a range of practical issues that arise in using developmental trajectories as a method to explore developmental disorders. Our own experience of using this technique has led us to produce a manifesto of rules-of-thumb, shown in Table 2. In this section, we briefly expand on two of the most important points: interpreting null results and dealing with variability.

Interpreting null results. In some of the examples described above, there were conditions where no reliable trajectory was found in the disorder group. That is, the function linking age and performance did not pick up a statistically significant amount of the variance. What does it mean when there is *no systematic relationship*? Does it really mean that performance does not improve with age in the disorder? This would be a pattern that radically departs from the expectations of normal development. However, although it could be true given the data, one has to take care with this interpretation. This is because *the cross-sectional design confounds differences in the severity of the disorder with differences in age.* Most disorders show a good deal of variability in how severely each individual is affected. When constructing a cross-sectional sample, there will not necessarily be a relationship between how severely each individual is impaired and how old they are (and indeed one hopes there will not be – to have, say, all your younger children more severely impaired than your older children would represent a recruitment bias). However, a decorrelation between severity and age means that any relationship between age and performance may be weakened or eliminated. It is nevertheless possible that if each individual were to be followed longitudinally he or she might show improvement, even while the cross-section as a whole does not. It is crucial to draw this distinction between cross-sectional and longitudinal designs. Most of the non-reliable trajectories of the preceding section occurred when a cross-sectional trajectory was constructed linking performance with chronological age. By contrast, in these same groups, reliable trajectories were found in the cross-sectional sample between performance and mental age. This is because severity is factored into the mental age: individuals who are more severely impaired will obviously have lower mental ages.

If a null result is found in a relationship between performance and mental age, there are several follow-up questions that must be asked. Were individuals in the disorder group able to understand the demands and carry out the test, given their level of ability? If one identifies a comparable task where the same group shows a reliable trajectory (as in the example of low-functioning children with autism), this increases confidence that the non-reliable trajectory of the first test is real. The influence of floor or ceiling effects may also destroy a relationship between performance and age; by contrast, if the disorder group scores in the sensitive range of the test, this also increases confidence that the non-reliable trajectory is real. Additionally, assuming the TD trajectory is satisfactorily linear, it may be that a non-linear trajectory is appropriate for the disorder and may predict a significant amount of the variability. Lastly, some trajectories can be reliable (statistically significant) but predict a very small amount of the variance, so that performance increases only slightly across the age range sampled. On the one hand, this is a benefit, because the trajectory design makes clear the difference between effect size and statistical significance. This distinction is sometimes omitted in the analysis-of-variance designs common in the matching methodology, where the important criterion is a significant difference between CA or MA control group to establish delay or deviance. On the other hand, one may legitimately ask what is a sufficient amount of variability for a trajectory to pick up before it should be taken seriously. For example, Figures 2 to 9 depict significant trajectories whose R^2 vary from .09 to .98. The answer to this question is that it depends on the effect size that one is expecting given the theory, given the experimental paradigm, and given the existing literature. The poorest fitting trajectories in our examples arose when performance was predicted by chronological

age rather than mental age, when performance was close to floor or ceiling reflecting limits on test sensitivity, and when more noisy reaction time data were used.

Given the likelihood that in many cases, the trajectory linking performance and chronological age for the disorder group will fall below that for the TD group, and given the problems of variations in severity destroying the relationship in cross-sectional analyses, one might ask why it is worth building CA-based trajectories for disorders – why not jump straight to considering developmental relations and hence trajectories constructed against MA? There are four reasons why CA-based trajectories are an important preliminary step in characterising a disorder. First, there will be abilities where we do not necessarily expect individuals with disorders to score more poorly (e.g., non-verbal skills in children with SLI). In these cases, the CA trajectories should coincide with the TD trajectory and be statistically different from the CA trajectories in areas of weakness. Second, CA trajectories are a theory-neutral description of how performance on average tends to improve with age in a disorder (subject to the limitations of cross-sectional designs). By contrast, MA-based trajectories are theory-driven. Third, by definition, the study of developmental relations focuses on relative abilities and this may mask absolute differences in comparisons to typical development. Thus it has been argued that in Williams syndrome, the developmental relation between Mean Length of Utterance and syntactic complexity is normal (i.e., not significantly different from the TD trajectory for this relation) and therefore that language development is normal in the disorder (in contrast to, say, Down syndrome, where syntactic complexity is lower than expected) (Mervis et al., 2000). However, it is all too easy to focus on the normality of the relations and ignore the absolute patterns that indicate that the most salient feature of language development in Williams syndrome is delayed onset (i.e., the WS CA-based

trajectory is significantly different to the TD trajectory in its intercept) and some suggestion of premature asymptote (Grant, Valian, & Karmiloff-Smith, 2001; Zukowski, 2001). Lastly, the comparison of CA- and MA-based trajectories is important to avoid being seduced by novel developmental relations in disorders. For example, let us say that two abilities, A and B, are correlated in a cross-sectional disorder sample but not in the TD sample (e.g., language and verbal memory ability in children with WS; see Meyer-Lindenberg et al., 2006). This could be because abilities A and B are causally related in the disorder but not in typical development. However, it could also occur because abilities A and B are both constrained by severity (a common causal factor) in the disorder, a factor that does not operate in the TD sample. For these reasons, then, we believe the study of developmental relations in disorders must be complemented by the initial construction of task-specific CA trajectories.

Variability: validating cross-sectional designs with longitudinal follow up. The examples thus far have focused on cross-sectional trajectories. One of the criticisms that can be made of cross-sectional trajectories is that the task-specific function linking performance and age may be the trajectory of *no single one* of the individuals (cf. Robinson, 1950). 'Average' development may not exist in the target group, in which case the only appropriate way to study developmental change is via longitudinal designs. In a sense, this is a criticism that afflicts the matching methodology also: the mean score of a group may not be the score of any individual within the group. The comparison of linear regression equations instead of group means does not make the issue of variability go away, and variability has always been a particular difficulty in the study of disorders (see, e.g., Thomas, 2003). But the criticism is correct in that the optimal developmental design is longitudinal. However,

longitudinal designs have disadvantages too. They are costly, place a burden on participants, suffer relatively high drop out rates, and produce long lags between the start of a project and the report of final results.

A more time efficient and cost efficient design begins by constructing a cross-sectional study and then uses longitudinal follow up of some or all of the participants to validate the trajectories predicted by the initial study. This design permits immediate reporting of provisional results, followed by validation of those results in a longitudinal design that is more tolerant of participant drop out. Such longitudinal follow up can also reveal limitations in the cross-sectional trajectories arising from shortcomings in test sensitivity such as floor effects. For example, Figure 9 depicts two cross-sectional trajectories for a sample of 28 children with Williams syndrome between the ages of 5;5 and 12;1, plotting test age on the BPVS (Dunn et al., 1997) and test age on the pattern construction subtest of the BAS-II (Elliott et al., 1996) against CA. These trajectories replicate a pattern often observed with WS, showing a marked disparity between the development of receptive vocabulary and visuospatial skills. Descriptively, the results indicate that receptive vocabulary has a delayed onset and is developing at only a marginally delayed rate, while pattern construction has the delayed onset but a more severely slowed rate. Some years after these data were collected, we revisited a small subset of 4 of these children, after a delay of between 27 and 49 months. The repeated measures are indicated in Figure 10 with unfilled symbols; thin lines link each follow-up measure to the first measure.

We can now evaluate whether the longitudinal trajectories of these four children fall within the confidence intervals of the initial cross-sectional trajectory. The results on vocabulary development are in the affirmative. The only individual who falls below the predicted trajectory also fell below it to begin with – this child

had a more delayed onset than average but the same rate. By contrast, the pattern construction findings demonstrate that the initial trajectory was incorrect. Two of the children who were at floor to begin with remained at floor, but the other two children showed increases in performance at a much faster rate than predicted; indeed, the rate was comparable to vocabulary development. The follow-up data suggest that the initial pattern construction trajectory mistakenly averaged together floor effects with real developmental improvement. Were the results of the latter two children with WS to be representative, the implication would be that the true delay is one that impacts mainly on onset, within the age range studied, and that the children with WS vary in the severity of their delays in onset. A more detailed consideration of the use of longitudinal trajectories to validate earlier cross-sectional findings for vocabulary development and pattern construction in Williams syndrome can be found in Jarrold, Baddeley, Hewes and Phillips (2001). In the current context, the more general lesson is that trajectories should only be built using scores that are above floor and below ceiling.

==========

Insert Figure 10 about here

==========

Discussion

We began by considering two contrasting theoretical positions on the origins of learning disability, the *developmental* and *difference* stances (Bennett-Gates & Zigler, 1998) and a more recent instantiation of this distinction in classifying individual cognitive abilities as *delayed* or *deviant / atypical* in developmental disorders. For language disorders, Rice et al. (2005) have argued that 'the contrast between delayed versus deviant aspects of language acquisition shows considerable promise in providing an overarching perspective on the ways in which language impairments can be manifest' (p.21). The idea of delay depends on identifying resemblances between the cognitive abilities of a disorder group and those of a younger typically developing group. In the course of this paper, we have argued that the opportunity to find these resemblances depends to some extent on the experimental methodology being employ. The use of developmental trajectories provides more ways in which similarities can occur between a disorder group and younger typically developing controls than the use of matched control groups. A richer descriptive vocabulary for characterising the ways that typical development can be deflected can only be helpful in seeking explanations for the impairments we observe in different disorders.

More widely, we view the strengths and limitations of the trajectory approach as follows. (1) Trajectories encourage researchers to place the developmental process at the heart of explanations of developmental deficits (Karmiloff-Smith, 1998). Although a methodology brings with it no necessary theoretical commitment, the requirement to derive a function characterising behavioural change over time focuses research in a way that can sometimes be lost when age (and therefore time) is factored out of the design, as is the case in matching. (2) Trajectories allow for flexible matching, offering multiple comparisons between the disorder group and a task-

specific typical developmental trajectory. Trajectories constructed against chronological age provide a more theory-neutral characterisation of a disorder. Trajectories constructed against mental age measures or other experimental tasks allow the researcher to explore developmental relations between abilities. (3) Trajectories can be descriptively powerful, as illustrated by the way that they discriminated between different forms of developmental delay. Finally, (4) while the easiest trajectories to construct are cross-sectional, validation by longitudinal follow-up provides an efficient and productive design. On the down side, the trajectories method relies on testing a wide age range of participants and the *availability of tests with sensitivity across that range*. Where the behaviour of interest is only found in a narrow age range, or tests have limited sensitivity, trajectories are not an optimal design and matching may be better. The study of behaviour across a wide age range also opens the trajectory approach to the criticism that there is no guarantee that behaviour on the same test is being driven by the same process at different ages. Indeed, there may even be a difference between the typical and disorder groups on the processes responsible for performance at different ages. This is an intrinsic problem in studying development and one that motivates an appeal to multiple converging sources of evidence, such as those provided by developmental cognitive neuroscience.

We illustrated the utility of trajectories by considering developmental delay but in finishing, we turn to the other possibility – that development in a disorder is *atypical*. What implications does the trajectory approach have for identifying qualitatively atypical developmental profiles? Delay corresponds to three types of relationship between the typically developing trajectory and the disorder trajectory where both generate reliable linear trajectories: delayed onset, delayed rate and delayed-onset+slowed-rate. These descriptions depend on the significance or non-

significance of differences in the intercept or gradient of regression lines. Atypicality (deviance, disruption) corresponds to the following possibilities: (1) although a reliable linear trajectory exists for the typically developing group, a non-linear function is a better fit for the disorder group, or there is no reliable trajectory for the disorder group. In the latter case, we distinguished between a zero trajectory and no systematic gradient. Particularly in longitudinal studies, a zero trajectory on an ability assessed with a sensitive measure implies a system that has reached its limit in undergoing ontogenetic changes. (2) Neither CA nor any theoretically relevant MA predicts performance in the disorder group while it does in the typically developing group. (3) A (potentially theoretically unexpected) measure of MA predicts performance in the disorder group but not in the typically developing group. Here one must take care that the novel association is not an artefact of variations in severity in the disorder but not the typically developing group in cross-sectional designs. Or (4) the same measures of MA predict performance to different extents in the typical and disorder groups

Note that the latter ascriptions of atypicality based on unexpected developmental relations appeal to an implicit mechanistic account where the cognitive system is taken to develop in blocks or domains (e.g., verbal, non-verbal, spatial) that causally affect each other. The lack of an expected MA predictor might indicate the absence of the block in the disorder, while the presence of an unexpected MA predictor might indicate atypical blocks or developmental contingencies. In each of these cases of atypicality, the markers of "qualitative" difference rely on (sometimes arbitrary) quantitative cut-offs, i.e., that a non-linear function gives a better fit than a linear function or that the relationship between a predictor and performance is significantly different between groups.

Importantly, the only non-quantitative way to identify deviance over delay in a disorder at the level of mechanism relies on the intuition of the experimenter in classifying errors. If a disorder group produce errors that are deemed qualitatively different based on the researcher's experience, a marker of atypicality is claimed (see, e.g., Scerif et al., 2004, Capirci et al., 1996; Karmiloff-Smith et al., 1997; Phillips, Jarrold, Baddeley, Grant, & Karmiloff-Smith, 2004; Thomas et al., 2006, for examples of using errors to test for atypical mechanism; Clahsen & Temple, 2003, and Tager-Flusberg, 2000, for claims that cases of qualitatively atypical development in disorders are rare).

Finally, although the descriptive power of trajectories has led us to distinguish different forms of delay, in our own research they have also led us to increasingly ask the following question. Why isn't delay itself atypical, especially when it is uneven across cognitive domains? One answer is that different rates of development are observed within normal development yet children generally reach the same endpoint (e.g., Bates et al., 1995; Dale et al., 2003). Delay in the normal range does not appear to be important. But is delay outside of the normal range the same kind of thing? These questions can only be addressed via mechanistic accounts, and may require the use of more formal neurocomputational models of development to properly specify the putative mechanisms causing developmental delay (Elman et al., 1996; Mareschal et al., 2007; Thomas et al., 2007). However, before we can engage on such an enterprise, we must have an accurate description of the target behavioural data to be explained by developmental models. And in this, we believe developmental trajectories provide a potent way forward.

References

Annaz, D. (2006). *The development of visuospatial processing in children with autism, Down syndrome, and Williams syndrome*. Unpublished PhD thesis. University of London.

Annaz, D., Karmiloff-Smith, A. & Thomas, M. S. C. (in press). The importance of tracing developmental trajectories for clinical child neuropsychology. In J. Reed & J. Warner Rogers (Eds.), *Child neuropsychology: Concepts, theory and practice*.

Annaz, D., Johnson, M. H., & Thomas, M. S. C. (2007). Contrasting cognitive profiles of 5-12 year old children with Williams syndrome, Down syndrome, high-functioning autism, and low-functioning autism: A comparison of developmental trajectories. *Manuscript in preparation*.

Ansari, D., Donlan, C., & Karmiloff-Smith, A. (in press). Typical and atypical development of visual estimation abilities. *Cortex*.

Bates, E., Dale, P. & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. pp. 96-151). Malden, MA: Blackwell Publishers

Bennett-Gates, D. & Zigler, E. (1998). Resolving the developmental-difference debate: An evaluation of the triarchic and systems theory models. In J. A. Burack, R. M. Hodapp, & E. Zigler (Eds.), *Handbook of mental retardation and development* (pp. 115-131). Cambridge: Cambridge University Press.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109-127). Oxford: Oxford University Press.

Bishop, D. V. M. (1997). Cognitive neuropsychology and developmental disorders: Uncomfortable bedfellows. *Quarterly Journal of Experimental Psychology, 50A*, 899-923.

Brock, J. (in press). Language abilities in Williams syndrome: a critical review. *Development and Psychopathology*

Brock, J., & Jarrold, C. (2004). Language influences on verbal short-term memory performance in Down syndrome: Item and order recognition. *Journal of Speech, Language, and Hearing Research, 47*, 1334-1346.

Brock, J., & Jarrold, C. (2005). Serial order reconstruction in Down syndrome: Evidence for a selective deficit in verbal short-term memory. *Journal of Child Psychology and Psychiatry, 46*, 304-316.

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 *Sixth International Professional Conference of the Williams Syndrome Association*, San Diego, CA.

Capirci, O., Sabbadini, L., & Volterra, V. (1996). Language development in Williams syndrome: A case study. *Cognitive Neuropsychology, 13*, 1017-1039.

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

Clahsen, H., & Temple, C. (2003). Words and rules in Williams syndrome. In Y. Levy, Y. & J. Schaeffer (Eds.), *Towards a definition of Specific Language Impairment in children*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97*, 332-361.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual route and parallel distributed processing approaches. *Psychological Review, 100*, 589-608.

Cook, R. D. (1977): Detection of influential observations in linear regression. *Technometrics, 19*, 15-18.

Cornish, K., Scerif, G., & Karmiloff-Smith, A. (in press). Tracing syndrome-specific trajectories of attention over the lifespan. *Cortex*.

Dale, P. S., Price, T. S., Bishop, D. V. M., & Plomin, R. (2003). Outcomes of early language delay: 1. Predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research*, 46, 544-560.

Dunn, Ll. M., Dunn, L.M., Whetton, C. & Burley, J. (1997). *British Picture Vocabulary Scale 2nd edition (BPVS-II)*. NFER-Nelson Publishing Company Limited: Windsor, Berks.

Elliott, C. D., Smith, P. & McCulloch, K. (1996). *British Ability Scales Second Edition (BAS II)*. NFER-Nelson Publishing Company Limited: Windsor, Berks.

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, Mass.: MIT Press.

Grant, J., Valian, V., & Karmiloff-Smith, A. (2002). A study of relative clauses in Williams syndrome. *Journal of Child Language, 29*, 403-416.

Hodapp, R.M., & Zigler, E. (1990). Applying the developmental perspective to individuals with Down syndrome. In D. Cicchetti & M. Beeghly (Eds.), *Children*

with Down syndrome: A developmental perspective (pp. 1-28). New York: Cambridge University Press

Hodapp, R. M., Burack, J. A., & Zigler, E. (Eds.) (1990). *Issues in the developmental approach to mental retardation.* New York: Cambridge University Press.

Jarrold, C., Baddeley, A. D., Hewes, A. K., Phillips C. (2001). A longitudinal assessment of diverging verbal and non-verbal abilities in the Williams syndrome phenotype. *Cortex, 37(3),* 423-431

Jarrold, C., Baddeley, A. D., & Phillips, C. (in press). Long-term memory for verbal and visual information in Down syndrome and Williams syndrome: Performance on the Doors and People Test. *Cortex.*

Jarrold, C. & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders, 34(1),* 81-86.

Jarrold, C., Cowan, N., Hewes, A. K., & Riby, D. M. (2004). Speech timing and verbal short-term memory: Evidence for contrasting deficits in Down syndrome and Williams syndrome. *Journal of Memory and Language, 51,* 365-380.

Karmiloff-Smith, A. (1997). Crucial differences between developmental cognitive neuroscience and adult neuropsychology. *Developmental Neuropsychology, 13 (4),* 513-524.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences, 2(10),* 389-398.

Karmiloff-Smith, A., Grant, J., Berthoud, I., Davies, M., Howlin, P. & Udwin, O. (1997). Language and Williams syndrome: How intact is “intact”? *Child Development, 68 (2),* 246-262.

Karmiloff-Smith, A., Scerif, G., & Ansari, D. (2003). Double dissociations in developmental disorders: Theoretically misconceived, empirically dubious. *Cortex*, *39*, 161-163.

Karmiloff-Smith, A., Thomas, M. S. C., Annaz, D., Humphreys, K., Ewing, S., Grice, S., Brace, N., Van Duuren, M., Pike, G., & Campbell, R. (2004). Exploring the Williams syndrome face processing debate: The importance of building developmental trajectories. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 45(7), 1258-1274.

Karmiloff-Smith, A., Tyler, L. K., Voice, K., Sims, K., Udwin, O., Howlin, P., & Davies, M. (1998). Linguistic dissociations in Williams syndrome: evaluating receptive syntax in on-line and off-line tasks. *Neuropsychologia*, *36*, 343-351.

Klein, B. P., & Mervis, C. B. (1999). Cognitive strengths and weaknesses of 9- and 10-year-olds with Williams syndrome or Down syndrome. *Developmental Neuropsychology*, *16*, 177-196.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *Autism Diagnostic Observation Schedule-WPS (ADOS-WPS)*. Los Angeles, CA: Western Psychological Services.

Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism Volume 1: How the brain constructs cognition*. Oxford: Oxford University Press.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders*, *49*, 333-348.

Mervis, C. B., & Robinson, B. F. (2003). Methodological issues in cross-group comparisons of language and cognitive development. In Y. Levy & J. Schaeffer

(Eds.), *Language competence across populations: Towards a definition of specific language impairment* (pp. 233-258). Mahwah, NJ: Lawrence Erlbaum Associates.

Mervis, C. B., Robinson, B. F., Bertrand, J., Morris, C. A., Klein-Tasman, B. P., & Armstrong, S.C. (2000). The Williams syndrome cognitive profile. *Brain & Cognition, 44*, 604-628.

Meyer-Lindenberg, A., Mervis, C. B., & Berman, K. F. (2006). Neural mechanisms in Williams syndrome: a unique window to genetic influences on social cognition and behaviour. *Nature Reviews Neuroscience, 7(5)*, 380-393.

Motulsky, H., & Christopoulos, A. (2004). *Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting.* Oxford: Oxford University Press.

Paterson, S. J., Brown, J. H., Gsödl, M. K., Johnson, M. H. & Karmiloff-Smith, A. (1999). Cognitive Modularity and Genetic Disorders. *Science, 286*, 2355-2358.

Phillips, C. E., Jarrold, C., Baddeley, A., Grant, J., & Karmiloff-Smith, A. (2004). Comprehension of spatial language terms in Williams syndrome: Evidence for an interaction between domains of strength and weakness. *Cortex, 40*, 85-101.

Rice, M. L. (2004). Growth models of developmental language disorders. In M. L. Rice & S. F. Waren (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 207-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Rice, M. L., Warren, S. F., & Betz, S. (2005). Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile X, specific language impairment, and Williams syndrome. *Applied Psycholinguistics, 26*, 7-27.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Rossen, M., Klima, E. S., Bellugi, U., Bihrle, A., & Jones, W. (1996). Interaction between language and cognition: Evidence from Williams syndrome. In J. H. Beitchman, N. Cohen, M. Konstantareas & R. Tannock (Eds.), *Language Learning and Behaviour* (pp. 367-392). New York, NY: Cambridge University Press.

Saemundsen, E., Magnusson, P., Smari, J., & Sigurdardottir, S. (2003). Autism Diagnostic Interview-Revised and the Childhood Autism Rating Scale: Convergence and discrepancy in diagnosing Autism. *Journal of Autism and Development Disorders, 33(3),* 319-329.

Scerif, G., Cornish, K., Wilding, J., Driver, J. & Karmiloff-Smith A. (2004). Visual selective attention in typically developing toddlers and toddlers with fragile X and Williams syndrome. *Developmental Science, 7,* 118-130.

Scerif, G., Karmiloff-Smith, A., Campos, R., Elsabbagh, M., Driver, J., & Cornish, K. (2005). To look or not to look? Typical and atypical development of oculomotor control. *Journal of Cognitive Neuroscience, 4,* 591-604.

Schopler, E., Reichler, R., & Rochen, B. (1993). *The Childhood Autism Rating Scale*. Western Psychological Services: Los Angeles.

Singer Harris, N. G., Bellugi, U., Bates, E., Jones, W., & Rossen, M. (1997). Contrasting profiles of language development in children with Williams and Down syndromes. *Developmental Neuropsychology, 13,* 345-370.

Strauss, M. E. (2001). Demonstrating specific cognitive deficits: A psychometric perspective. *Journal of Abnormal Psychology, 110,* 6-14.

Tager-Flusberg, H. (2000). Differences between neurodevelopmental disorders and acquired lesions. *Developmental Science*, *3*, 33-34.

Tager-Flusberg, H., Plesa-Skwerer, D., Faja, S., & Joseph, R., M. (2003). People with Williams syndrome process faces holistically. *Cognition, 89*, 11–24.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thomas, M. S. C. (2003). Multiple causality in developmental disorders: Methodological implications from computational modelling. *Developmental Science, 6(5)*, 537-556.

Thomas, M. S. C. (2007). Language acquisition in developmental disorders. In M. Kail, M. Hickmann & M. Fayol (Eds.), *Proceedings of the International Conference on First and Second Language Acquisition*, Paris 2006.

Thomas, M. S. C., Dockrell, J. E., Messer, D., Parmigiani, C., Ansari, D., & Karmiloff-Smith, A. (2006). Speeded naming, frequency and the development of the lexicon in Williams syndrome. *Language and Cognitive Processes, 21(6)*, 721-759.

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. (2002). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioural and Brain Sciences, 25(6)*, 727-780.

Thomas, M. S. C. & Karmiloff-Smith, A. (2003). Modelling language acquisition in atypical phenotypes. *Psychological Review, 110(4)*, 647-682.

Thomas, M. S. C., Karmiloff-Smith, A., Scerif, G., Ansari, D., & Johnson, M. H. (2007). Is the concept of 'delay' explanatory in the study of developmental disorders? Theoretical and neurocomputational considerations. *Manuscript in preparation*.

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.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review, 98*, 3-53.

Yule, W. (1978). Research methodology: What are the "correct controls"? In M. Rutter & E. Schopler (Eds.), *Autism: A reappraisal of concepts and treatment* (pp. 155-162). New York: Plenum.

Zukowski, A. (2001). *Uncovering grammatical competence in children with Williams syndrome*. Unpublished doctoral thesis, Boston University.

Figure captions

Figure 1: example of data from matching design

Figure 2: example of data from a developmental trajectory design

Figure 3: simulated data of (a) reliable and (b) non-reliable trajectories showing improvements in performance with age on a target ability (y-axis uses an arbitrary scale; (c) two non-reliable trajectories with different variance around the regression line; (d) the two trajectories after 45˚ anti-clockwise geometric rotation – only the trajectory with small variance becomes reliable.

Figure 4: the shape of delayed (a-c) and atypical (d-e) developmental trajectories, along with the same data plotted in terms of group means (f) on an experimental task (y-axis scale is arbitrary).

Figure 5: past tense elicitation performance for TD and WS groups, for regular (talk) and irregular (drink) verbs, plotted against CA. MA was measured using the BPVS (Dunn et al., 1997). (Data from Thomas et al., 2001).

Figure 6: naming times for numerals 1-9 for TD and WS groups plotted against CA (log-log transformed) (data from Thomas et al., 2006). MA was measured using the BPVS (Dunn et al., 1997).

Figure 7: Comparison of test age scores for the BPVS (Dunn et al., 1997) and the pattern construction sub-test from the BAS-II (Elliott et al., 1996) plotted against CA, for TD, high-functioning children with autism, and low-functioning children with autism (data from Annaz et al., 2007).

Figure 8: Non-linear developmental trajectories for verbal recall and recognition tests (data from Jarrold et al., in press). Verbal mental age (VMA) was measured using the BPVS (Dunn et al., 1997).

Figure 9: Verbal recall and recognition performance standardised for chronological or verbal mental age according to the BPVS (Dunn et al., 1997). (Data from Jarrold et al., in press).

Figure 10. Comparison of test age scores for 28 children with WS on the BPVS (Dunn et al., 1997) and the pattern construction sub-test of the BAS-II (Elliott et al., 1996) plotted against CA. Unfilled symbols show longitudinal follow-up scores for 4 of the children, within thin lines illustrating individual longitudinal trajectories.

Figures

Figure 1

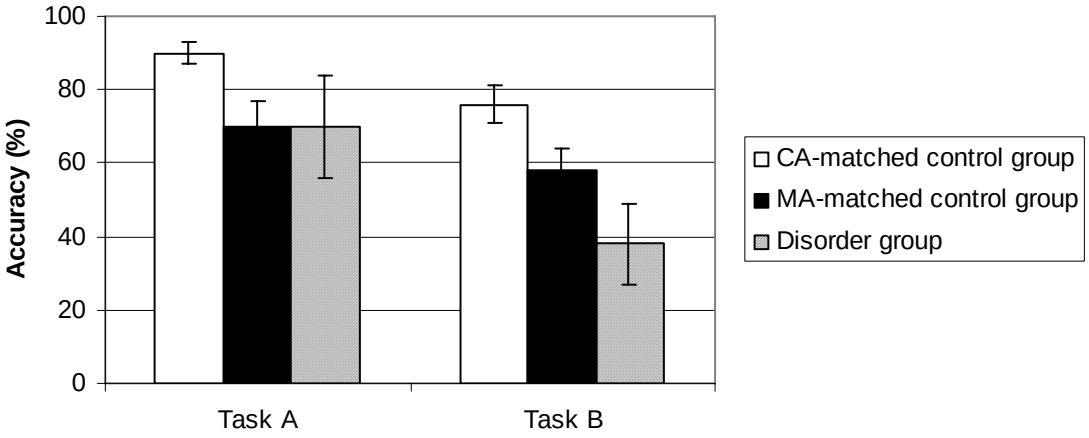
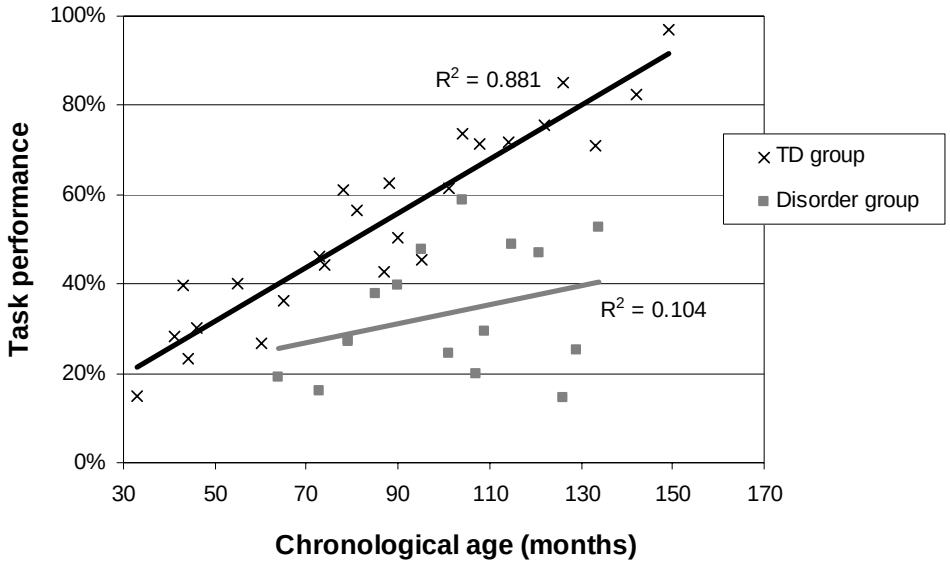


Figure 2

(a)


100%
80%
60%
40%
20%
0%
30
50
70
90
110
130
150
170
Task performance
Chronological age (months)
R2 = 0.881
R2 = 0.104
TD group
Disorder group

(b)

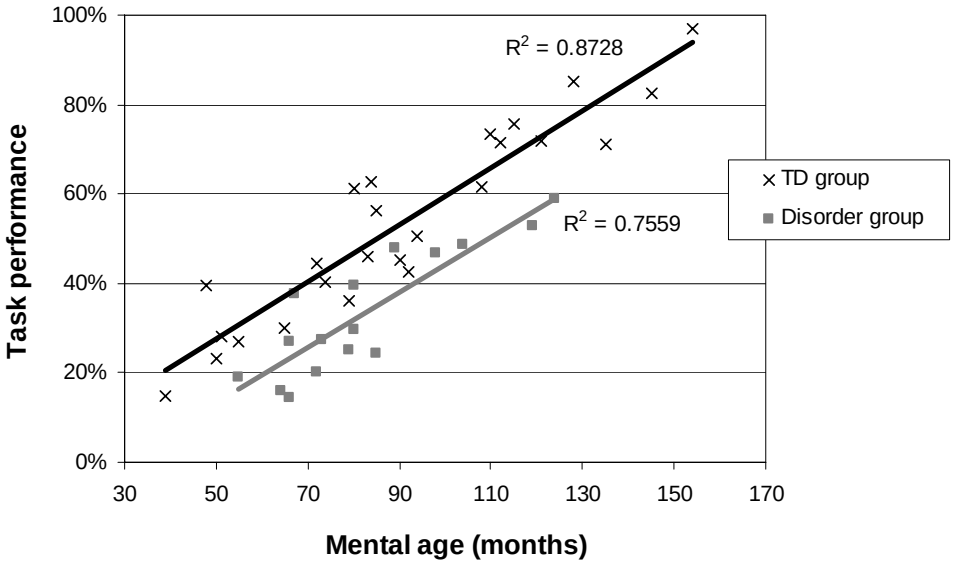
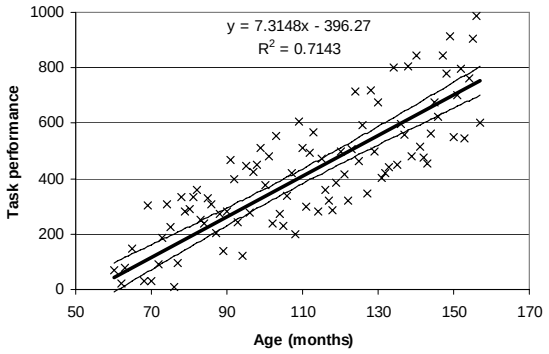
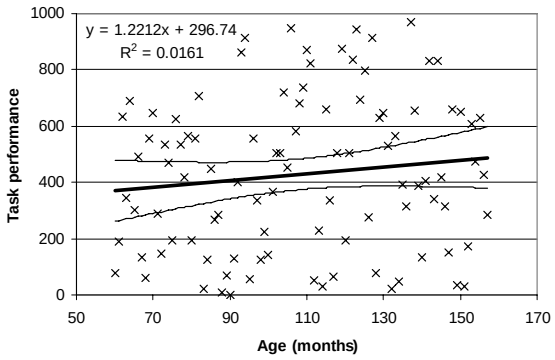

100%
80%
60%
40%
20%
0%
30
50
70
90
110
130
150
170
Task performance
Mental age (months)
R2 = 0.8728
R2 = 0.7559
TD group
Disorder group

Figure 3

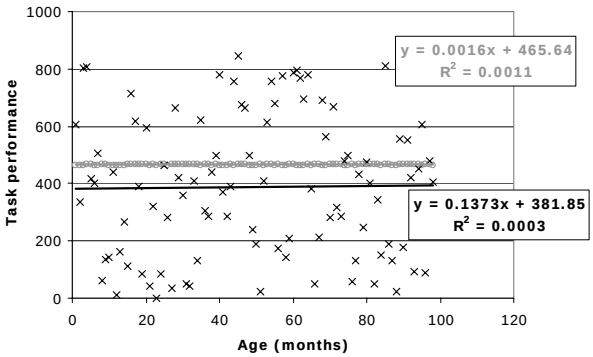
(a)



(b)



(c)



(d)

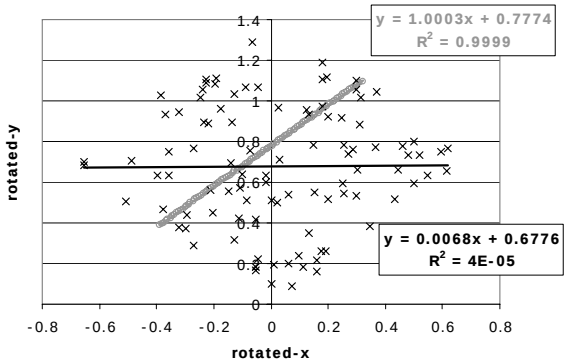


Figure 4

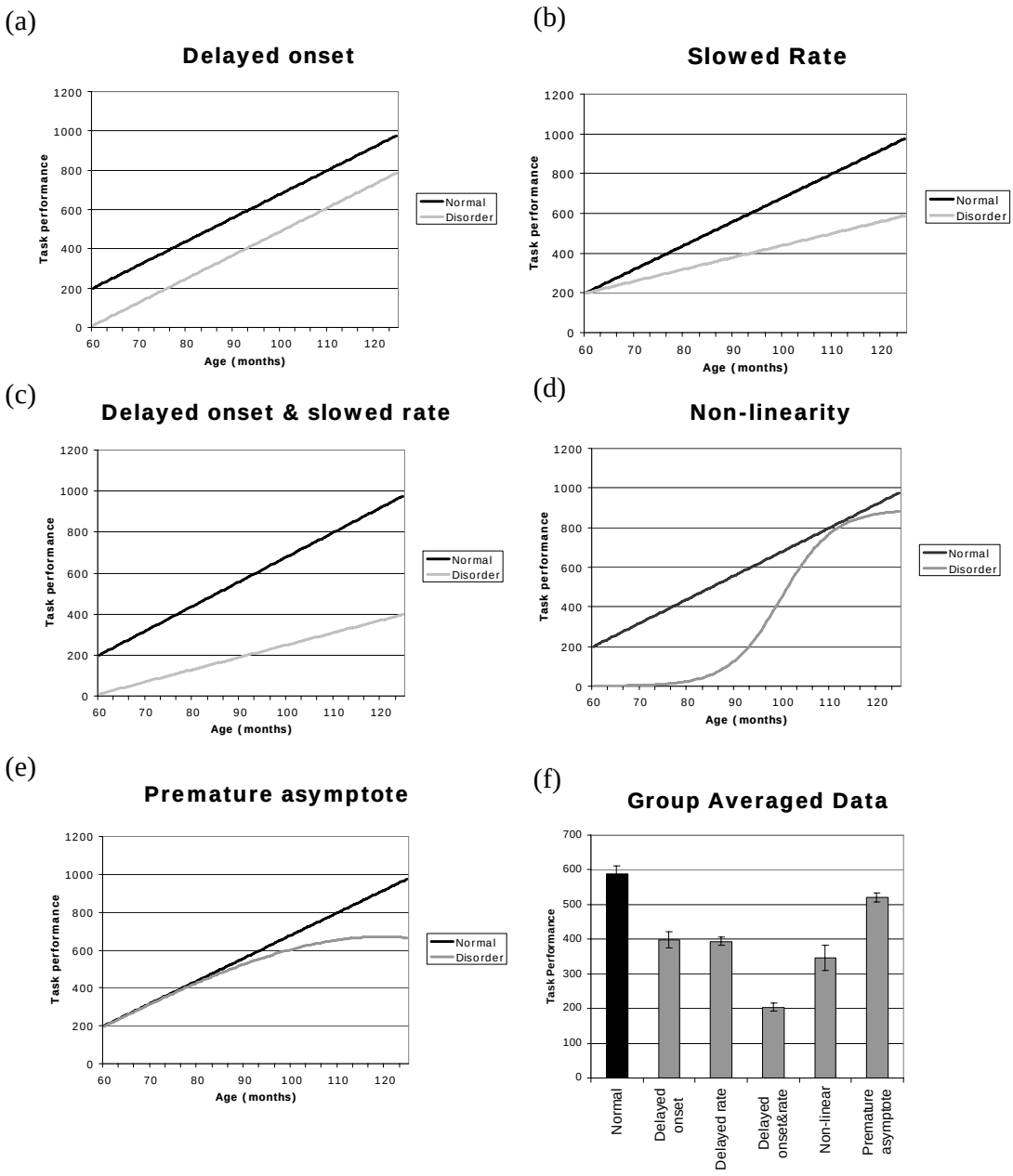

(a)
Delayed onset
Task performance
Age (months)
Normal
Disorder
(b)
Slowed Rate
Task performance
Age (months)
Normal
Disorder
(c)
Delayed onset & slowed rate
Task performance
Age (months)
Normal
Disorder
(d)
Non-linearity
Task performance
Age (months)
Normal
Disorder
(e)
Premature asymptote
Task performance
Age (months)
Normal
Disorder
(f)
Group Averaged Data
Task Performance
Normal
Delayed onset
Delayed rate
Delayed onset&rate
Non-linear
Premature asymptote

Figure 5

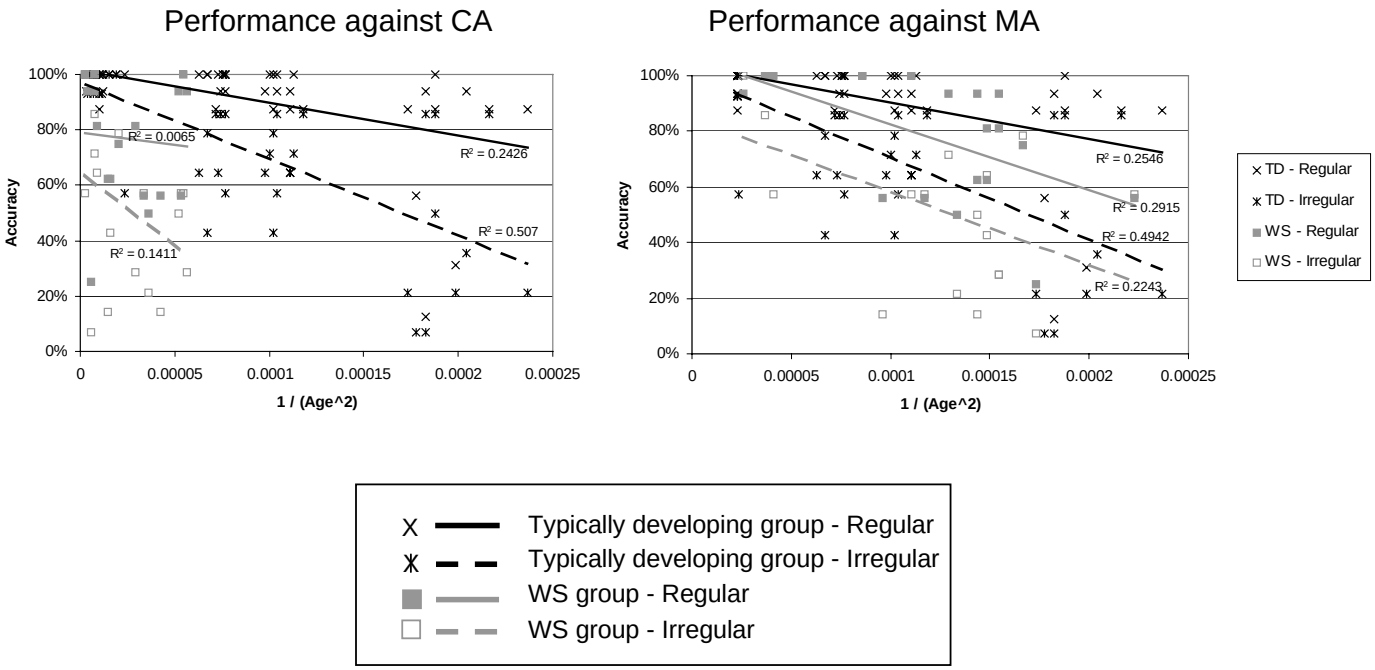

Performance against CA
Performance against MA
Accuracy
100%
80%
60%
40%
20%
0%
0
0.00005
0.0001
0.00015
0.0002
0.00025
1 / (Age^2)
R² = 0.0065
R² = 0.2426
R² = 0.507
R² = 0.1411
R² = 0.2546
R² = 0.2915
R² = 0.4942
R² = 0.2243
TD - Regular
TD - Irregular
WS - Regular
WS - Irregular
Typically developing group - Regular
Typically developing group - Irregular
WS group - Regular
WS group - Irregular

Figure 6

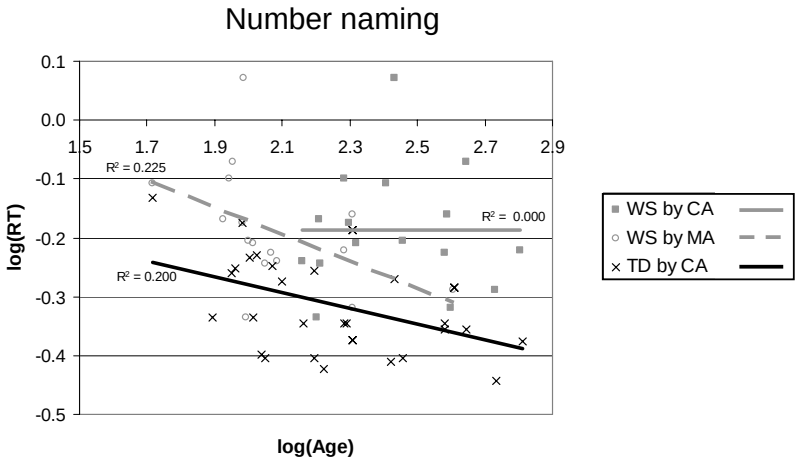

Number naming
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
1.5
1.7
1.9
2.1
2.3
2.5
2.7
2.9
log(RT)
log(Age)
R² = 0.225
R² = 0.000
R² = 0.200
WS by CA
WS by MA
TD by CA

Figure 7

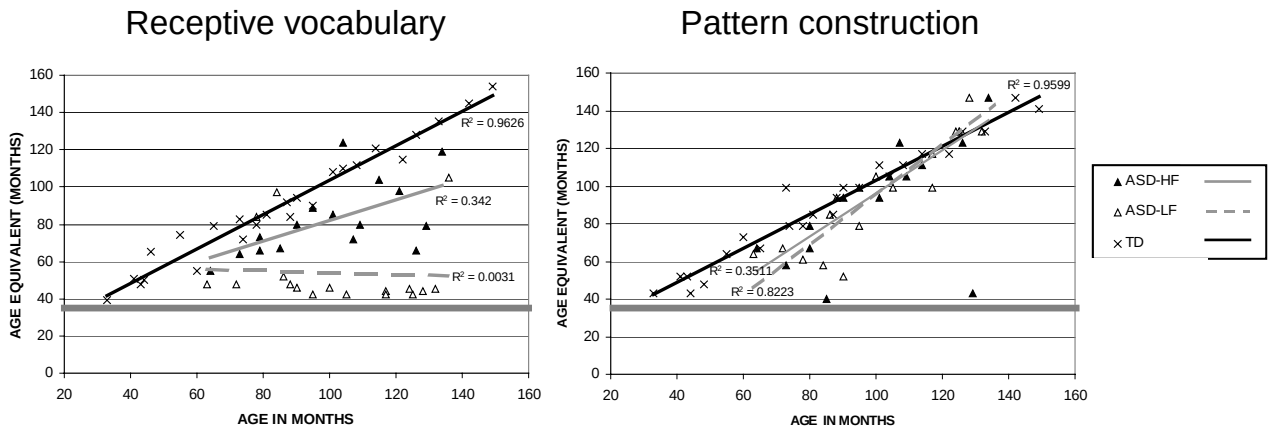

Receptive vocabulary
Pattern construction
AGE EQUIVALENT (MONTHS)
AGE IN MONTHS
AGE EQUIVALENT (MONTHS)
AGE IN MONTHS
R² = 0.9626
R² = 0.342
R² = 0.0031
R² = 0.9599
R² = 0.3511
R² = 0.8223
ASD-HF
ASD-LF
TD

Figure 8.

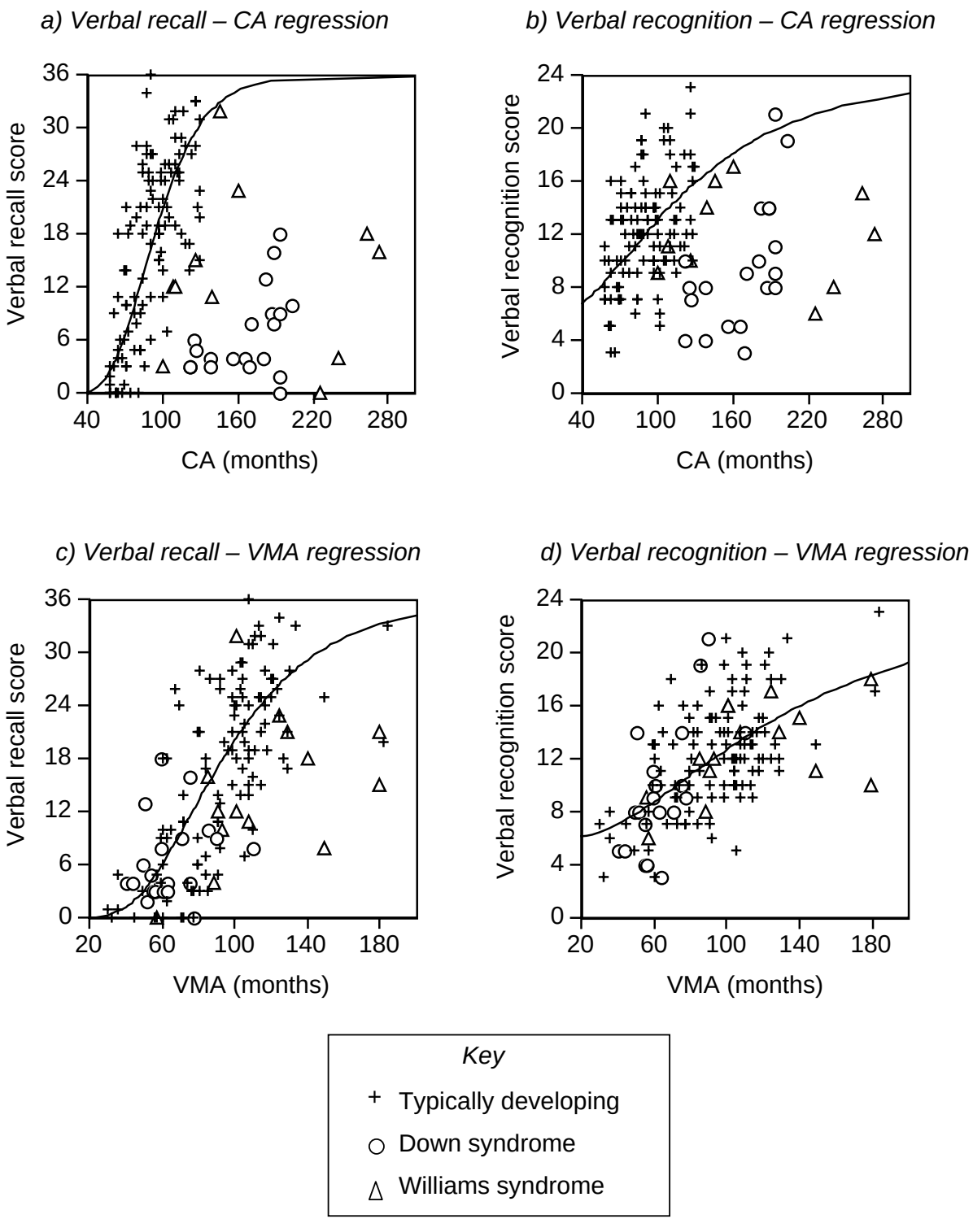

a) Verbal recall – CA regression
b) Verbal recognition – CA regression
c) Verbal recall – VMA regression
d) Verbal recognition – VMA regression
Verbal recall score
Verbal recognition score
CA (months)
VMA (months)
0
6
12
18
24
30
36
4
8
16
20
40
100
160
220
280
20
60
140
180
Key
Typically developing
Down syndrome
Williams syndrome

Figure 9

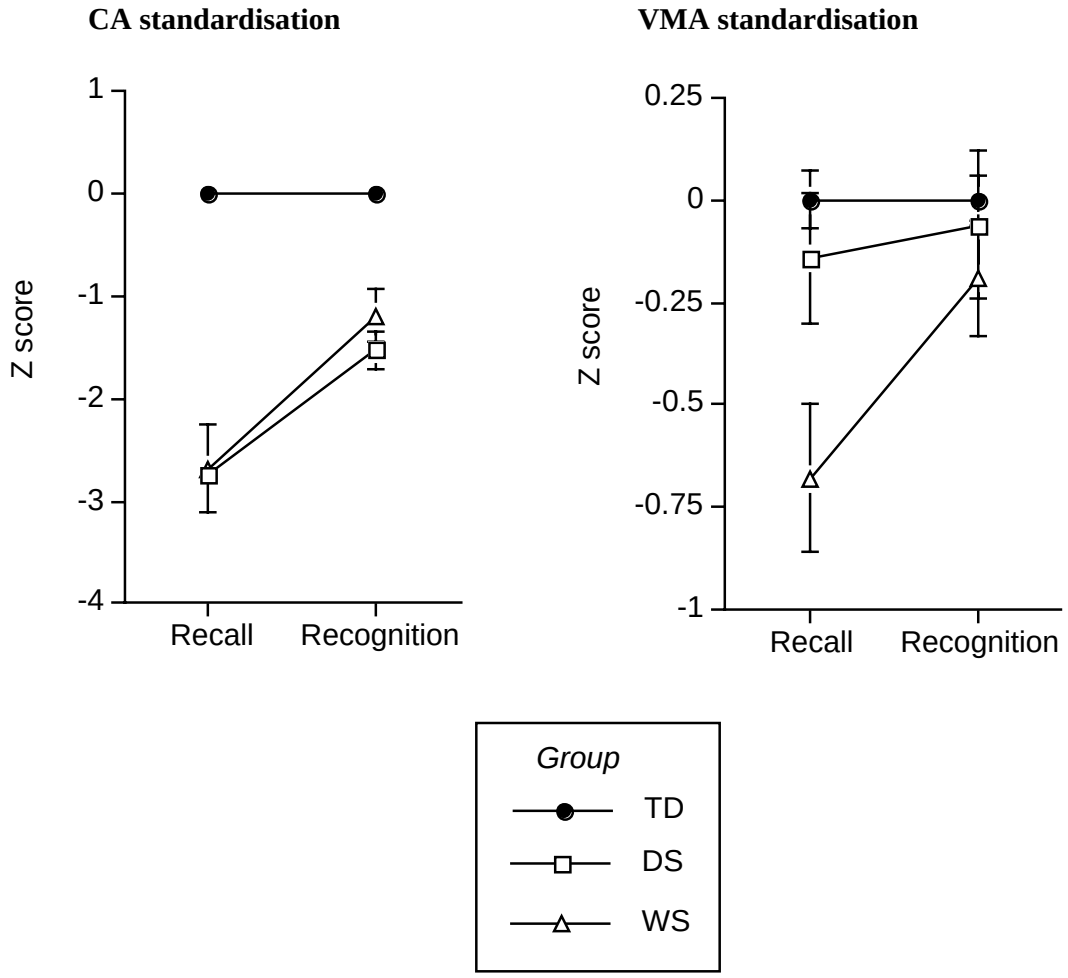

CA standardisation
VMA standardisation
Z score
1
0
-1
-2
-3
-4
0.25
-0.25
-0.5
-0.75
Recall
Recognition
Group
TD
DS
WS

Figure 10

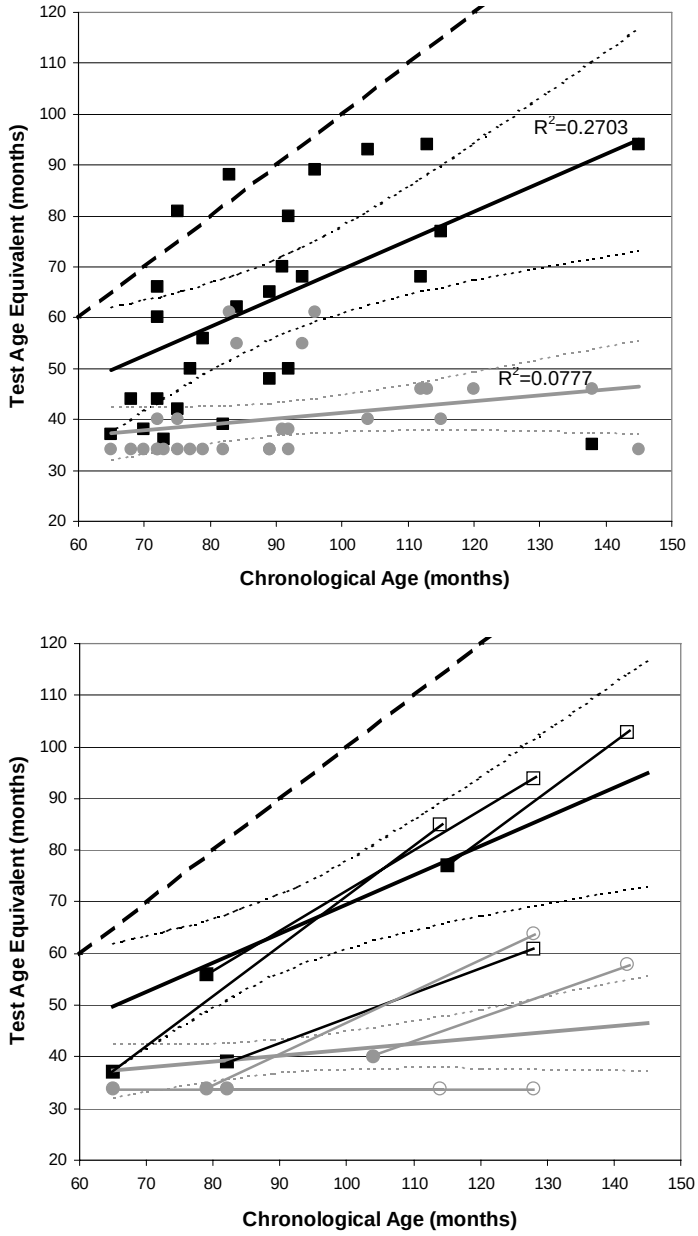

Test Age Equivalent (months)
Chronological Age (months)
R2=0.2703
R2=0.0777
CA-level expected score
WS BPVS test age equivalent
BPVS longitudinal follow-up
WS PC test age equivalent
PC longitudinal follow-up

Table 1: Comparison of the methodologies for investigating developmental disorders

Methodology	Individual matching	Developmental Trajectories
Age range	Narrow age range	Wide age range
Comparisons	Chronological-age matched control group	Theory-neutral ("can each individual from the disorder group fit *anywhere* on the TD trajectory?")
	Theory-dependent mental-age match (1 control group per theory)	Performance predicted by chronological age
		Performance predicted by mental age of disorder group (as many comparisons as standardised tests run on disorder group) or by performance on other experimental tasks to derive *developmental relations*
Discrimination	In sensitive range of test (can be narrow) Avoid floor and ceiling effects	In sensitive range of test (must be wide) Avoid floor and ceiling effects
Aim	Factor out age from comparison	Derive function relating performance to age

Table 2. Rules-of-thumb for trajectory analyses

- Try not to use the descriptor “delay” on its own
    - If it is intended simply to mean “not at the level predicted by CA”, then this phrase should be used instead
- In the statistical characterisation of trajectories, delay terms are only intended to be descriptive
    - No mechanism is implied
- Test should be in sensitive range across trajectory where possible
    - Minimally, the test should be sensitive across the ability range of the disorder group
    - If there are ceiling or floor effects, trajectories can be constructed just for the age range above floor or below ceiling (if there are sufficient participant numbers)
- TD and disorder groups should be evenly distributed across the age range where possible
    - Trajectories cannot be compared across non-overlapping age ranges
- Use linear methods for interpretability
    - Transform data to linear if necessary
    - Ensure that the trajectory accounts for sufficient amount of variability
    - Outliers should not have undue influence on the regression line (use e.g., Cook’s distance, 1977)
        - If outlier has undue influence, then report trajectory with and without outlier; seek to explain outlier
    - Regression line should not radically violate assumptions of linearity and homogeneity of variance (standardised residuals should be normally distributed, plot of standardised predicted scores against standardised residuals should show no pattern)
    - In diagrams, always depict variability where practical: plot all data points and label each trajectory with its R^2 value (percentage variance account for by the trajectory); for completeness, report gradients and intercepts with their confidence intervals as well
    - Confidence intervals around regression lines can be used to measure the convergence or divergence of trajectories (CI=region within which there is 90% chance that mean lies)
- Types of delay
    - Delayed onset (significant difference in regression line intercepts)
        - Delays in onset should be calculated from the regression equation for the minimum age / mental age measured in the TD group
    - Slowed rate (significant difference in regression line gradients)
    - Delayed onset + slowed rate (significant difference in intercepts and gradients)
- Additional descriptors
    - Non-linearity
    - Premature plateau
- Developmental relations
    - Is task performance predicted by mental age measures derived from standardised tests relevant to the target domain (according to current theory)?
    - Is task performance predicted by performance on other experimental measures?
- Validate cross-sectional developmental trajectories with longitudinal follow-up where possible