

Modelling Change: New Opportunities in the Analysis of Microgenetic Data

Andrea Cheshire^{a,*}, Kevin P. Muldoon^b, Brian Francis^c,
Charlie N. Lewis^a and Linden J. Ball^a

^aDepartment of Psychology, Lancaster University, Lancaster LA1 4YF, UK

^bHeriot-Watt University, UK

^cDepartment of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK

Despite the increasing use of the microgenetic methodology to examine change, the techniques employed to analyse microgenetic data remain fairly unsophisticated. This paper reviews the existing ways of analysing such data and describes their limitations. We use two recent studies to illustrate how modelling can avoid these problems and reveal important aspects of children's cognitive development. The first example illustrates the use of quasi-binomial modelling to examine 6- and 7-year olds' analogical reasoning development. This method offered insights into the way in which children develop, in terms of the rate and path of change, and how different instructional cues can affect their performance. The second study employs a random effects logistic model to analyse the development of preschoolers' counting skills. This technique was employed to examine different influences on children's use of counting to compare quantities. We argue that the key benefit of such modelling approaches is that they are able to tap into the process of change whilst not compromising statistical assumptions. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: microgenetic; modelling; cognitive development; change

If microgenetic studies are superior to longitudinal and cross-sectional designs it is because the intensive nature of data can reveal detailed information about what actually occurs during periods of rapid change. Rather than inferring the processes that lie between the initial and end states of a transition, these methods make it possible to analyse change from a number of different perspectives. However, such goals can only be achieved if researchers have the analytic framework and tools for assessing the process of change. In this paper, we argue

*Correspondence to: Andrea Cheshire, Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK. E-mail: a.cheshire@lancaster.ac.uk

that microgenetic research could be even more innovative if data analysis matched the complexity of the designs employed.

In the first half of the paper we review the analytic approaches taken in recent microgenetic research. While the evidence has been compelling and the range of approaches to analysis has been broad, we argue for an approach that attempts further to reveal the determinants and processes of change. In the second half we describe, in brief, a modelling approach to the analysis of microgenetic data that aims to tackle the assessment of critical change. In order to support our case we then explore two complementary databases to show how development can be modelled. We aim to generate discussion about the analysis of change and demonstrate how practical solutions can be achieved.

TENSIONS IN THE APPROACHES TO ANALYSING MICROGENETIC DATA

By design, microgenetic experiments involve intensive data collection. Notwithstanding the general consensus that the more data one collects, the greater the potential to discover important associations that might otherwise have been missed, there are clear statistical issues that need to be addressed. In order to examine change, microgenetic datasets often include several conditions, multiple sessions and a high number of trials or observations per session. Of essential interest is the nature of *transitions*, involving the path, rate, breadth, variability and sources of change (Siegler, 1995, 1996). These five themes are often analysed, to some extent, in most microgenetic studies. However, the majority of analytical techniques employed still leave the actual developmental processes to be inferred. Most studies continue to rely on 'traditional' analysis of variance (ANOVA) approaches to identify change by examining general patterns in the variability within and between individuals over time (e.g., Blöte, Resing, Mazem, & van Noort 1999; Blöte van Otterloo, Stevenson, & Veenman 2004; Siegler, 1995; Taylor & Cox, 1997; Tunteler & Resing, 2002). One benefit of ANOVA is that it is sensitive to both the shared patterns of performance of individuals across time and to variability between groups. Importantly, it is also accepted within psychology as the standard approach to group comparisons in both cross-sectional and longitudinal designs. Repeated measures ANOVA additionally allows the effects of individual factors to be assessed and variance sources (such as those identifying between-individual and within-individual variation) to be separated and identified. A skilled analyst can specify effect contrasts that enable particular treatment effects to be tested.

However, for microgenetic data, the use of ANOVA is problematic. First, it relies upon assumptions of consistent variance within and between individuals, yet such assumptions concerning homogeneity of variance may be breached in a number of ways. The nature of many microgenetic experiments is to collect test results that are naturally discrete (usually identifying the shift from failing to passing on trials of a particular measure), and individuals approaching the floor or the ceiling of such tests will commonly have a lower variance than those returning a central or mid-point value. With discrete data, the assumptions of normality that are needed for ANOVA are typically breached. Moreover, the design used may not be balanced—there may be differing numbers of measurements at different times for each individual. Finally, the focus of ANOVA is on variance decomposition rather than parameter estimation, and complex microgenetic hypotheses concerning change are best expressed through para-

metric model testing rather than through the comparison of variances explained; we return to this point below.

Non-parametric methods which attempt simply to provide a distribution-free alternative to ANOVA do not provide a way forward. Their emphasis is on simple hypotheses and commonly used tests are unable to take into account random effects, complex model specifications and other dependence structures. Techniques using ranks also do not provide the flexibility needed to examine hypotheses relating to change.

ALTERNATIVE APPROACHES

In order to overcome the problems encountered using analyses that rely upon blanket comparisons of groups over time, a number of other approaches have been used. These tend to share a focus on individual patterns of change in order to tease apart the processes involved and the variations between individuals. Three such approaches are outlined here.

1. *Case study analysis*: Whereas ANOVA examines differences between groups, the case study approach takes the opposite stance, examining each individual participant's data in depth. Such case-by-case analysis is believed to provide insight into the features of the process of interest (Kuhn & Phelps, 1982). The large amount of data can reveal much about the dynamics of change and detailed individual accounts of how people reach the same level of performance.

Although this approach typically examines data from each participant separately, Lavelli, Pantoja, Hsu, Messinger, and Fogel (2005) acknowledge further possibilities. By combining each case study they claim that it is possible to examine group differences and 'isolate a particularly interesting developmental transition in order to study how individuals, dyads, or groups navigate across that transition' (p. 49). However, they do not make clear how this kind of data would be analysed, leaving the reader to assume that one would revert to traditional approaches of analysing differences between groups. We conclude that case study analyses can reveal important details about development such as the trajectories of change and individual differences but, if the aim is to infer common pathways and general patterns of influence, case study analyses cannot be used in isolation.

2. *Graphical methods*: Backward trials graphing (Opfer & Siegler, 2004; Siegler, 1996; Siegler & Svetina, 2002) has been used to examine the development of skills and conceptual understanding. This method can be used to complement more traditional ways of analysing data such as ANOVA and non-parametric equivalents. The aim of this technique is to find what influences the individual's performance prior to a specific strategy change taking place. This is achieved by producing a series of graphs that analyse the path and variability of change.

First, a '0 trial block' is created that marks the earliest point at which each child uses a strategy consistently. For example, Siegler and Svetina (2002) employed a criterion of three consecutive trials using a strategy of interest to designate a '0 trial block'. The point that children reach this level of consistency during a study is likely to be different for each child. Thus one participant's '0' point might occur on the 10th trial of the experiment, another on the 20th, and another on the 30th (Siegler, 2006). The blocks of trials preceding and following this '0' point are calculated separately for each child and are labelled -2 , -1 and $+1$, $+2$ and so on. These blocks of trials indicate the percentage of children using particular strategies before and after their discovery of the strategy of interest.

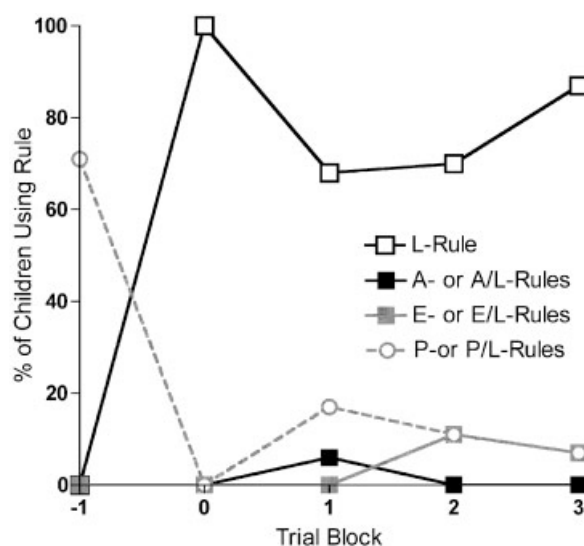


Figure 1. Example backward trials graph (Opfer & Siegler, 2004). These data points are estimates from the original graph.

In a recent study Opfer and Siegler (2004) employed a backward trials graphing method to explore their findings on categorizing biological concepts. They created four graphs, with the '0 trial block' representing the point when children first identified that all living things (i.e. plants and animals) should be categorized as (1) living, (2) being able to reach a goal, (3) being able to grow, and (4) needing water to live. Figure 1 illustrates the rules that children used to categorize living things when they were asked whether or not each item grows. Here, Opfer and Siegler highlight the use of four different rules that the children employed, with the key interest being when children categorize all living things as being able to grow. The figure illustrates how the predominant but incorrect 'P- or P/L rule' (only plants, or some animals and all plants, are categorized as being able to grow) decline as the correct 'L-rule' (all living things are categorized as being able to grow) becomes dominant. This method is also particularly insightful as it illustrates the pattern of rule usage after a correct rule has been discovered.

The types of pattern identified in Figure 1 illustrate the utility of this technique in the way it provides important information about when a change takes place and what strategies were employed immediately before and after discovery of the correct rule. However, like the case study approach, the backward trial graphing technique is primarily focused upon analyses of a single group. It cannot tell us whether there are significant differences between experimental manipulations. This is not a damning critique of the approach as the microgenetic method concerns the concentration of data collection and could include, for example, observations of naturally occurring phenomena. However, our main criticism is that backward trials graphing is a visual approach that does not subject the data to hypothesis testing: 'visual analysis is notorious for its subjectivity, unreliability, and insensitivity in detecting less pronounced changes' (Lavelli et al., 2005, p. 53). So, backward trial graphing is highly informative but may be accused of falling short of the demands of sufficiency in terms of methodological rigor.

3. *Fuzzy sets*: The aim of fuzzy sets analysis is to provide a more accurate and informative characterisation of participants' performance than arises from simply analysing mean scores. Van Geert (2002) highlights the importance of examining the variability in human performance, commenting that in any given situation scores not only vary between people and between conditions, they also vary within individuals. He points out that the properties of a particular task, any one participant, or a specific context are not clear cut, and that it is more useful to consider these properties as having dynamic and fuzzy boundaries.

When examining an individual participant's performance over repeated sessions it is possible to detect a range of scores on very similar trials of any given task. Van Geert stresses the importance of examining this variability in order to understand development and claims that it is possible to work out how characteristic an individual's performance is of his/her ability at any given point in time. He uses 'fuzzy logic' (see van Geert, 2002, for a more detailed description) to quantify how typical an observed score is in terms of the task and context of testing. This approach allows specification of the path of development, thereby enabling analysis of the changing distribution of participants' scores over the period of testing. Van Geert also claims that being able to specify the width of the range of performance allows for more accurate predictions of future performance.

The main benefit of fuzzy sets analysis is being able to take into consideration the entire range of scores that a participant produces over the course of testing. This can reveal interesting and useful information about the development of particular skills in a particular context. The limitations of this method are similar to those of the case study approach. Fuzzy sets analysis tends to concentrate on individual participants, making it difficult to examine any experimental manipulations and group differences. This is not to imply that all microgenetic research examines differences between groups of children. As illustrated above, some studies focus on the trajectories and fluctuations of an individual participant's performance. However, we contend that statistical analyses of microgenetic data should allow for the analysis of group data.

MODELLING MICROGENETIC DATA

Most statistical procedures can be represented through the concepts of statistical modelling and the primary aim of this paper is to demonstrate that such models provide a key ingredient to the analysis of microgenetic data that complement the more descriptive approaches taken above. Statistical models, when specified correctly, allow researchers to investigate specific hypotheses and should be used with great caution if they are employed in hypothesis-generation. The goal of the modeller is to build a statistical model for a variable of interest (the response variable) which adequately represents the type of the variable and the data being collected. Such data might reveal shared patterns of change or different trajectories between subgroups, including patterns like U-shaped developmental functions that have been associated with some developmental changes (such as preschoolers' learning of non-verbal symbols, Namy, Campbell, & Tomasello, 2004, and the use of determiners in language development, Karmiloff-Smith, 1979). These developmental trajectories typically show a period of intense change, which then slows down and reaches a new plateau (see Siegler, 2004). The modelling approach illustrated in this paper attempts to identify the nature of these changes and should also describe the way in which other variables act on

the measure of interest and the nature of any dependence relations (see Moskowitz & Hershberger, 2002, for an introduction to modelling repeated measures designs). The primary aim is to separate out structural effects from noise and random error. To do this the model would be defined by parameters representing particular aspects of the underlying theoretical position to be explored. The parameters are estimated from the data by maximizing the likelihood function, which represents the likelihood of the observed data given the distributional assumptions made. We flesh out an example of this process below.

In considering simple statistical models for cross-sectional data, we would need to determine a suitable parametric distribution for the response variable. The nature of the response needs to be considered, with different distributions needed for continuous data, count data, and proportions. With longitudinal data we have a number of additional key ingredients that would feed into the specification of the statistical model:

- (a) As we are following the same participants through time, what dependence structure do we wish to specify? Can we assume that there is an individual-specific or random effect, which is the same for all observations from the same person?
- (b) What is the measurement structure over time? Is the same number of measures taken from participants at exactly the same points in time? Or are the data more observational in nature, with different numbers of follow-up times and different time points?
- (c) Does the observation at the current time point depend on measurements made *at* the current time point, or is there also a lag effect, with dependence on variables measured at different time points?
- (d) How does the treatment regime change over time, and how is it best to represent this in the statistical model? Do we need, for example, to consider change-point models where the regression slopes change at particular time locations?

The specification of a statistical model for longitudinal data is a complex exercise, and texts by Agresti, Booth, Hobert, and Caffo (2000) and McCulloch and Searle (2000) provide thorough descriptions of the methodology. Software to fit such models is still limited. We use GLIM4 as it allows models to be clearly specified (e.g. Francis, Green, & Payne, 1993) but other statistical packages such as R and STATA also give the modelling flexibility required. Statistical models are becoming increasingly common in texts on longitudinal data analysis (Singer & Willett, 2003). These models are not uniformly in line with our approach. For example, growth curve modelling (e.g. Jordan, Hanich, & Kaplan, 2003) allows for individual random effects and assumes quadratic effects over time. Such models have, however, typically assumed normality for the underlying distribution of the test responses, and do not deal with floor or ceiling effects.

The approach we are proposing here is a more flexible specification of the data structure, both through specification of the underlying distribution, and in terms of the non-linear relation of change over time. This allows us to move away from the constraining assumptions of a quadratic curve (in which steep growth is followed by a plateau then followed by a decline). Non-normal random effects models are now commonly used in medicine, biology and psychiatry, but are less commonly used in psychology. They are useful as they extend the classic normal random effects model to other types of data (count data, binary data, ordinal

data) where the assumption of normality for the response is no longer appropriate (see Agresti *et al.*, 2000). Additionally, visualisation of fitted models will play an important part, which cannot be used in techniques like structural equation modelling.

We proceed by considering two recent examples. These share the common goal of attempting to model differences between experimental groups by comparing the nature of change within individuals. The first example employs quasi-binomial distribution random-effects models to analyse the influence of different types of feedback on the development of analogical reasoning skills. The second examines how data can be analysed using a random-effects logistic regression model to explore the development of the precursor skills to the acquisition of children's ability to use counting rather than length as a basis of set comparisons.

Example 1: Self-explanation and the development of analogical reasoning skills

This example illustrates how quasi-binomial distribution models can be used to analyse microgenetic data. In this study (Cheshire, Ball, & Lewis, 2005) we examined the development of children's analogical reasoning skills, focusing on the impact of self-explanation (i.e. asking children to provide a rationale behind their selected response to a given task) and feedback (i.e. informing children if their response was correct, and, if not, then showing them the correct answer). This study was based upon the microgenetic approach of Siegler and Svetina (2002). We extended their experiment by designing five conditions in which children experienced different combinations of explanation and feedback. Children in all but a control group took part in seven sessions. Those in the control group only took part in sessions 1 and 7 to examine any natural development over that time period. The first six sessions took place at 3–5 day intervals with the seventh session taking part approximately 8 weeks later, with the aim of the last session to be a measure of stability of learning. In each session participants completed 22 trials of graded difficulty to test analogical reasoning skills. In the first session, to gain an initial level of competence, children were presented with a reasoning task with no explanation or feedback requirements. In sessions 2–5 children were instructed according to the explanation and feedback requirements of the condition. In sessions 6 and 7 children were no longer required to provide explanations nor were they given feedback, so that the constraints on children were similar to that in session 1 and were therefore comparable.

When a sample is repeatedly tested there will invariably be wide variations both between and within individuals. Figure 2 illustrates this variability in one of the groups that were tested in this study. Although this variance can be of interest (i.e. even when given the same treatment why do some children make progress whilst others do not?) the focus of the analysis in this paper is to examine the difference in performance at the group level. When the data of participants in each group are aggregated and compared they produce patterns that smooth out individual differences and suggest variations between groups. In Figure 3, performance of each group is displayed across the time-frame of the study. It highlights differential effects of training, but the key question concerns how best to analyse this dataset. Applying ANOVA to these data reveal main effects of time and group, plus interactions between the two. However, this analysis could not take into consideration the control group who only participated at two time points and, more importantly, such data fall foul of many issues discussed in the first section of this paper (e.g. breaching assumptions concerning homogeneous variance).

Figure 3 also provides information about the potential issues that may be worth exploring using modelling techniques, particularly as the main hypotheses concerned the relative advantage of children receiving both feedback and being

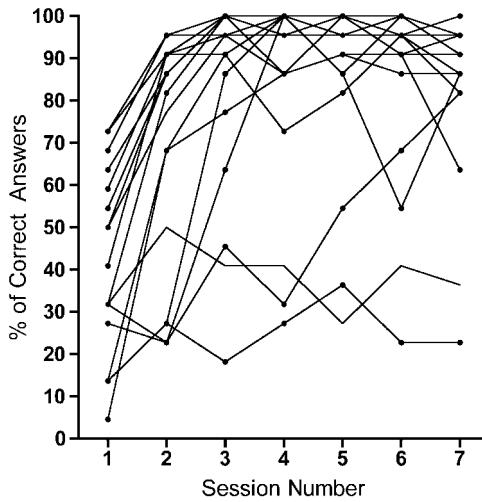


Figure 2. Individual trajectories of observed data in the 'explanation plus feedback' group.

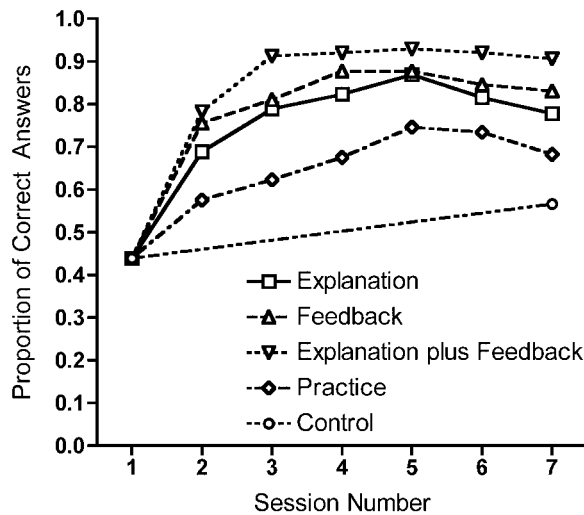


Figure 3. Mean probability of children in each group selecting a correct response in each sessions.

required to explain their judgements. By setting out the patterns of performance across an appropriate time frame, Figure 3 suggests that the nature and rates of change differ between groups and differentially at each time point. We wished to test differences in the rate of change and to see if the dip between the last training point and the first post-test was important. The main issue concerns the relative rate of children's performance over the seven microgenetic sessions in terms of group membership. To examine this, the proportion of correct answers (i.e. the number of successes out of the total number of trials) for each individual, per session, was entered into a model.

The proportion of correct answers for each individual per session is usually correctly modelled as arising from a binomial distribution, where the probability of a correct answer would be analysed as a function of session, treatment and individual-level effects. However, in these data this principle is suspect as it assumes that all trials are of equal difficulty, whereas in fact some were potentially easier than others. This could inflate the variance of the binomial distribution, producing data that are overdispersed. One straightforward method for dealing with this is to inflate the variance by a multiplicative factor of k , which is estimated through the ratio of the chi-squared statistic to the residual degrees of freedom for the most complex model under consideration (Firth, 1991). This is akin to estimating the residual variance from the ratio of the error sum of squares to the residual degrees of freedom in analysis of variance.

To model the data, we need to consider both group changes over time, and individual-specific variability (or random effects), which take account of differential performance between individuals. We relate the proportion (p) of items correct to time and other explanatory variables through a logistic link function, or transformation $\log(p/1-p)$ which is a standard link function for such data (see Collett, 1991; Hosmer & Lemeshow, 1989). This transformation turns a probability (restricted to between 0 and 1) into a number that can take any value. Change over time is modelled flexibly, allowing for discontinuation in slope at various time points to take account of varying experimental conditions over time.

To address the crucial issue of the nature and rate of change in each group over time we first attempted to identify the important periods of change within the dataset. To do this we measured how many distinct change points across time were required, using theoretical issues within the literature as a basis (for example, Siegler and Svetina, 2002, Figure 3, p. 800). Figure 3 shows a decline in performance between the end of training (session 5) and the first post-test (session 6), but the analysis suggested that this was non-significant. The best model to fit the data was one with two change points: a rapid period of development up to 7 days, a slower rate of change up to the end of training, followed by maintenance of this level of performance (see Figure 4). Models with and without such dip showed no difference, so we can use the rule of parsimony to remove the dip.

A second key question concerns whether the rate of change is similar between groups across these three phases of change. The first way to address this question is to combine the data from different groups to see if this influences the overall model. If it does not then we can assume that the groups are similar in their overall profiles of change. The aim is to find the most parsimonious model with only the groups that truly differ from each other being distinguished. For example, it was possible to combine the 'explanation only' and 'feedback only' groups without a significant change in the deviance values (see Figure 5). However, it was not possible to combine all four experimental groups because the change in deviance scores was significant. In other words, this model did not appropriately reflect the data, and it is therefore possible to conclude there were significant differences between, at least, some of the experimental groups.

This approach, thus, allowed us to identify how and when change occurred in children's performance, and the way in which this performance differed between groups. It enabled us to conclude with confidence that children who improved most were those who generated explanations and were also given feedback. It also enabled us to deliberate between each of these factors administered individually. The joy of microgenetic data concerns their ability to allow detailed insight into many aspects of developmental change. It would be possible to

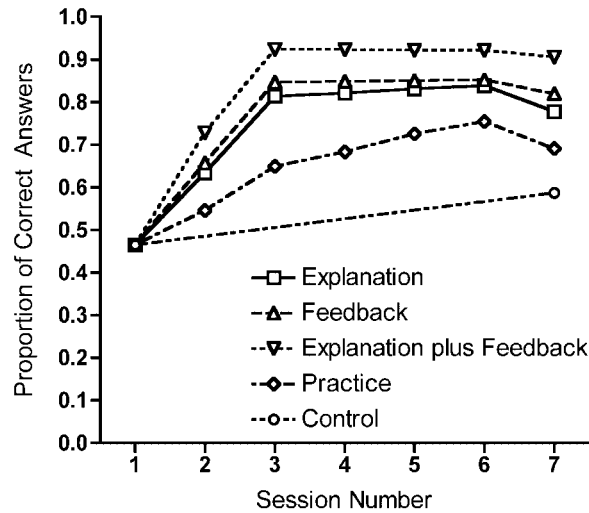


Figure 4. Statistical model with equal slopes up to session 3 and change points at session 3 and session 6.

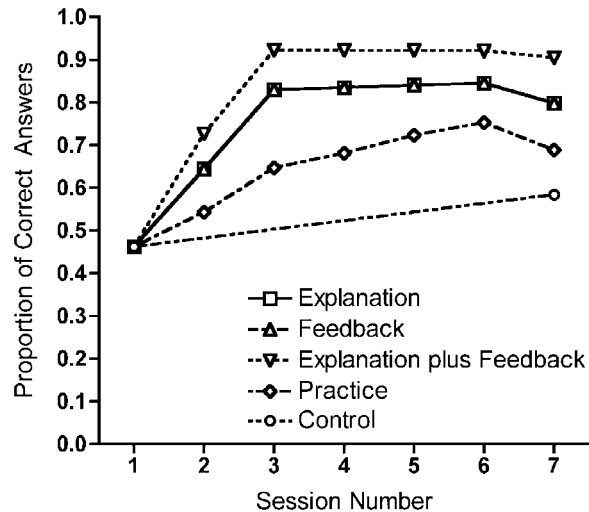


Figure 5. Simplified statistical model combining groups 1 and 2.

model further the factors surrounding the lack of change in some individuals or discrepancies in the general patterns displayed in Figure 2 and also to examine the data in terms of transitions between individual trials within each session.

Example 2: Early numeracy and set comparisons

A key question in early cognitive development is how children make transitions between simple skills, like counting a single set to produce a cardinal number, to more complex skills, like judging the quantitative relations between two or more sets by comparing the respective cardinals. One obstacle to this development appears to be that other cues to relative quantity, like spatial

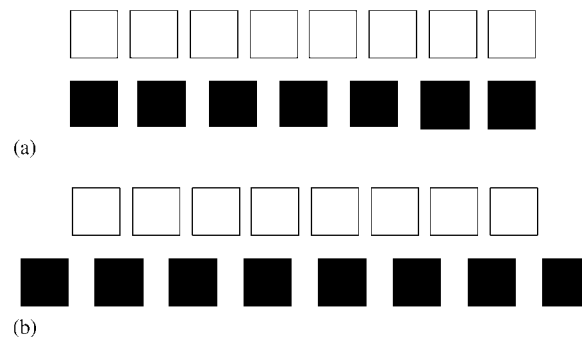


Figure 6. Set comparison task: (a) unequal number/equal length (UN/EL) problem; and (b) equal number/unequal length (EN/UL) problem.

magnitude, may actually mislead children. We began by identifying 4- and 5-year-olds who, although they could count accurately, consistently compared sets of coloured blocks (arranged in parallel rows such that length was always a salient but misleading cue) on the basis of their relative lengths rather than by counting them. During three training sessions, children were asked to judge similarly arranged sets of blocks, but this time they were counted, giving children the opportunity to compare cardinals instead of length. An important dimension to the training was that on half the trials, one of the rows was miscounted (by a puppet). The aim was to see whether children would (1) recognize the significance of the cardinal numbers for the task of comparing sets, and (2) appreciate the impact that miscounting has on cardinality.

The design was intended to assess social influences on children's learning. Following Siegler (1995), Group 1 were given feedback on whether their judgment was correct or not, but not asked to explain that judgment; Group 2 were asked, following their judgment, 'How did you know that?'; and Group 3 first received feedback and were then asked to explain the experimenter's verdict. The research question concerned whether any of these training regimes influenced children's ability to compare two rows that were either (1) unequal in number of items but equal in length (UN/EL), or (2) equal in number of items but unequal in length (EN/UL) (see Figure 6). A post-test measured children's spontaneous use of counting to compare similar arrays.

This design involves within-participant factors of Counting Accuracy (whether both rows were counted accurately or one row was miscounted), Problem-type (UN/EL versus EN/UL), Time (training sessions 1–3), and also the between participants effect of Group Membership. The problem was that analysing the data entailed using a score out of 2 (the number of trials for each Counting Accuracy \times Problem-type combination on each session for each Group) as the dependent measure. With a response set with only 3 possible values (0, 1 or 2), small sample sizes, and a hypothesis regarding change (which implicitly breaches the assumption of homogeneous variance), it was not plausible to assume an underlying normal distribution. As before, we used a mixed-effects logistic model to analyse the data for main effects and any interactions using the appropriate distributional assumption (i.e. binomial).

We used a macro in GLIM4 (Francis, *et al.*, 1993) called ALLVC, implementing a technique developed by Aitkin and Alfo (1998). This can fit a model where the response distribution is assumed to be binomial and the individual-specific

random effect is taken to be normally distributed. The dependent variable was thus declared to be the number of correct judgments, the error structure was taken to be binomial denoted by $B(n, p)$, (where n is the number of trials and p is the probability of passing any one trial) with the parameter n set to 2 (the number of trials for each combination of Counting Accuracy and Problem-type), and the link function was the logit.

The model was defined as follows: 'Group', 'Counting Accuracy', 'Problem-type', and 'Time' had a small number of experimental levels and were determined as part of the design, and thus these variables were entered as fixed-effects. A 'Participant' effect was included as a normally distributed random effect to allow for participants having repeated measurements across the three time phases of the study.

The aim of the statistical analysis was to identify which variables influenced learning during training *and* children's recognition that the skills applied there were also relevant on later tasks where instructions to count were not given. We used scores on training sessions 2 and 3, and the post-test as the dependent measures, with measures from training sessions 1, 2 and 3 as predictors, respectively (i.e. measures from training session 1 were regressed on to scores from training session 2 to see which, if any, predicted success). Thus the model is, essentially, three separate logistic regressions, with variables from Time X predicting children's scores at Time $X + 1$, and with separate sets of predictors being allowed for each time point. The random 'Participants' effect acknowledged the fact that each child was tested repeatedly and, therefore, each child had an individual effect on changes in performance.

One set of variables entered into the regressions accounted for mean changes associated with group membership on different sessions (e.g. membership of Group 1 during training session 1, repeated for sessions 2 and 3 and then again for Groups 2 and 3). The second set of variables considered how each group performed on particular types of array in each count condition during each training session (e.g. Group 1, session 1, on Accurate-Count EN/UL problems, repeated to cover all the interactions between Group, Time, Problem-type and Counting Accuracy). This gave a total of 45 new independent variables. Many of these variables needed to be included in the model as main effects, but of primary interest were the interactions. With 45 observations we ran the risk of over-fitting, but the focus of this analysis was to test whether the pre-specified interactions were significant. We loaded these variables, plus age, into the analysis, with the 'Participants' random effect also included to account for individual differences in performance. It was important to simplify the model, to remove unimportant terms, and to increase the accuracy of estimation of remaining parameters. There are numerous methods of achieving this, either using information criteria such as the Akaike Information Criterion (AIC), or by examination of changes in scaled deviance ($-2 \log$ -likelihood) when terms are removed. To avoid over-fitting, these methods must be combined with tests of hypotheses which derive from the existing literature. Additionally, full replication is recommended and in this series we replicated by comparing data from the microgenetic investigation with those from a similar longitudinal study. The patterns of data were consistent in the two studies.

The resulting model (see Table 1) reveals how success (i.e. correctly identifying when two sets are equal ('fair') or not ('unfair')) on training sessions 2 and 3, and the post-test, can be predicted by performance on the preceding session. The values in the right-hand column represent the odds ratio that a unit-increase on that variable has on a child's success on the subsequent session/post-test. For

Table 1. Predictors from session X of success on session $X + 1$

Predictor variable	Odds-ratio
<i>(a) Predictors from session 1 of success on session 2</i>	
Group 1—EN-UL—puppet counts accurately	4.05
Group 1—EN-UL—puppet miscounts	2.33
Group 2—EN-UL—puppet counts accurately	3.92
Group 2—EN-UL—puppet miscounts	2.01
Group 3—EN-UL—puppet counts accurately	5.33
Group 3—EN-UL—puppet miscounts	2.13
<i>(b) Predictors from session 2 of success on session 3</i>	
Group 1—EN-UL—puppet counts accurately	4.21
Group 2—EN-UL—puppet counts accurately	6.38
Membership of Group 3	8.01
<i>(c) Predictors from session 3 of success on Post-test</i>	
Membership of Group 1	-2.72
Group 1—EN-UL—puppet miscounts	2.27
Group 2—EN-UL—puppet miscounts	4.75
Group 3—EN-UL—puppet miscounts	5.67

Note: Success in sessions 2 and 3 entails being able to judge correctly whether a puppet has made the sets 'fair'. Success of the post-test entails spontaneously counting rows before comparing cardinals in order to judge fairness.

EN = equal number; EL = equal length; UL = unequal length.

example, the value 4.05 means that if a child from Group 1 makes a correct judgment on EN/UL trials during training session 1 when both rows are counted accurately, the odds of that child making a correct judgment during session 2 roughly quadruple.

The predictors from sessions 1 and 2 suggest that children in each group were benefiting from their successful reliance on cardinal numbers on 'Accurate-Count' trials, and applying this knowledge on the subsequent session. However, comparing the odds-ratio values highlights how group membership influenced the way children responded to the different problems. For example, on session 1, if a child judged the EN/UL problems successfully when both rows were counted accurately, their chances of making a correct judgment on session 2 increased roughly by a factor of four for children if they were a member of either Group 1 or 2, and by a factor of just over five if they were in Group 3. The important thing to note here is that in the absence of any main Group effect (i.e. none of the predictor variables representing membership of either Group 1, 2 or 3 at Time 1 are retained), the model reveals that if a child was able to make a correct judgment on these particular trials their chances of making a correct judgment on the next session increased significantly. Indeed, the retention of variables during backwards elimination was based on a relatively conservative criterion; only odds ratios greater than 2 (i.e., only those predictors that at least doubled the chances of scoring on the next session) were retained in the model.

Performance on 'Miscount' trials was rarely the most significant predictor of success during training. However, it was success on 'Miscount' trials during session 3 that predicted whether children would spontaneously use a counting strategy to compare rows on the post-test. Such findings show how children learned the importance of cardinal values (i.e. that they denote relative quantity) during training by recognizing their primacy over misleading spatial cues to

quantity. The results also highlight how children who were able to recognize and reflect on the consequences of miscounting were more likely to change their strategy from pre- to post-test. Importantly, children do not appear to benefit to the same extent from observing and being able to explain the significance of accurate counting strategies. Furthermore, the greater use of counting by Groups 2 and 3 on the post-test reveals the impact that social mediation (i.e. asking children to explain numerical strategies and judgements) can have in promoting this type of learning. The model also reveals how group membership alone sometimes predicted whether success was likely (e.g. membership of Group 3 at training session 2) or unlikely (e.g. membership of Group 1 at training session 3).

Modelling performance in this way can be a useful diagnostic tool when addressing limitations in children's mathematical ability. This approach to microgenetic data has the potential to reveal the effects of success or failure of a particular strategy on behaviour at subsequent time points (Miller & Coyle, 1999). Modelling these microgenetic data allowed us to see how children's reactions to feedback, their ability to explain particular reasoning, their conceptualisation of the counting principles, and communicative interaction with an adult promoted the use of counting as a problem-solving tool.

CONCLUSION

In this paper, we suggest that previous approaches to analysing microgenetic datasets have not always accounted for their complex nature. We have argued that the blanket application of analysis of variance techniques might not be the ideal solution to examining issues of developmental change as its underlying assumptions are in conflict with those of the microgenetic method. Alternative approaches like case studies certainly help to highlight variability in the nature and rates of change. Using a mixture of analytical approaches can also afford advantages for deriving a more penetrating understanding of transitions in cognitive development. We argue, however, that a statistical modelling approach can be a particularly valuable adjunct in such a mixed-method framework as modelling is able to tap into the process of change within and between groups whilst not violating statistical assumptions. Statistical modelling can take into consideration repeated testing by including a random effects element, which accounts for the variation in participants' performance over the period of time in which they are being tested. The models can be based on pre-specified statistical distributions of particular datasets (e.g. a binomial distribution) rather than on an assumption that the data are normally distributed.

Other statistical models can be considered. One of the most promising areas of future development is latent class modelling (see for example, Clogg, 1995); these models can estimate different 'trajectories' for different groups of participants in a study. Over sufficient data points, these statistical models can be used to model such individual trajectories. This provides a link between the fuzzy sets approach of van Geert (2002) and the approach described here. Another promising line of attack is graphical chain modelling (Edwards, 2000). One potential problem of the approach that we have advocated here is that it is possible to construct complex models involving as many variables and interactions as there are participants. It is important to use the techniques we have described here to test existing theoretical claims rather than to fish for spurious associations. Nevertheless, if used wisely, a modelling approach allows for greater confidence in the interpretation of data as the analysis can be pursued without having to violate

basic statistical assumptions. The challenge that remains is to develop these techniques in such a way that they are attractive and easy to use.

ACKNOWLEDGEMENTS

This research was funded by the Economic and Social Research Council (PTA-030-2003-00102 and PTA-026-27-0800). The statistical analyses were also supported by the ESRC (RES-576-25-5020). The authors wish to thank John Opfer who gave us permission to include his backward trials graphs in Figure 1. We are also grateful to three anonymous reviewers for their valuable comments.

REFERENCES

- Agresti, A., Booth, J., Hobert, J. P., & Caffo, B. (2000). Random effects modelling of categorical response data. *Sociological Methodology*, *30*, 27–80.
- Aitkin, M., & Alfo, M. (1998). Regression models for binary longitudinal responses. *Statistics and Computing*, *8*, 289–307.
- Blöte, A. W., Resing, W. C. M., Mazer, P., & van Noort, D. A. (1999). Young children's organisational strategies on a same-different task: A microgenetic study and a training study. *Journal of Experimental Child Psychology*, *74*, 21–43.
- Blöte, A. W., van Otterloo, S. G., Stevenson, C. E., & Veenman, M. V. J. (2004). Discovery and maintenance of many-to-one counting strategy in 4-year-olds: A microgenetic study. *British Journal of Developmental Psychology*, *22*, 83–102.
- Cheshire, A., Ball, L. J., & Lewis, C. N. (2005). Self-explanation, feedback and the development of analogical reasoning skills: Microgenetic evidence for a metacognitive processing account. In B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the twenty-seventh annual conference of the cognitive science society* (pp. 435–441). Mahwah, NJ: Lawrence Erlbaum Associates.
- Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modelling for social and behavioural sciences* (pp. 311–359). New York: Plenum.
- Collett, D. (1991). *Modelling binary data*. London: Chapman & Hall.
- Edwards, D. (2000). *Introduction to graphical modelling* (2nd ed.). New York: Springer.
- Firth, D. (1991). Generalized linear models. In D. V. Hinkley, N. Reid, & E. J. Snell (Eds.), *Statistical theory and modelling* (pp. 55–82). London: Chapman & Hall.
- Francis, B., Green, M., & Payne, C. (Eds.). (1993). *The GLIM system release 4 manual*. Oxford: Oxford University Press.
- Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.
- Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, *74*, 834–850.
- Karmiloff-Smith, A. (1979). *A functional approach to child language: A study of determiners and reference*. New York: Cambridge University Press.
- Kuhn, D., & Phelps, E. (1982). The development of problem-solving strategies. *Advances in Child Development*, *17*, 1–44.
- Lavelli, M., Pantoja, A. P. F., Hsu, H., Messinger, D., & Fogel, A. (2005). Using microgenetic designs to study change processes. In D. M. Teti (Ed.), *Handbook of research methods in developmental science* (pp. 40–65). Oxford: Blackwell.
- McCulloch, C. E., & Searle, S. R. (2000). *Generalized, Linear, and Mixed Models*. New York: Wiley.
- Miller, P. H., & Coyle, T. R. (1999). Developmental change: Lessons from microgenesis. In E. K. Scholnick, K. Nelson, S. A. Gelman, & Miller, P. H. (Eds.), *Conceptual development: Piaget's legacy* (pp. 209–239). London: Lawrence Erlbaum Associates.

- Moskowitz, D. S., & Hershberger, S. L. (Eds.). (2002). *Modelling intraindividual variability with repeated measures data*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Namy, L. L., Campbell, A. L., & Tomasello, M. (2004). The changing role of iconicity in non-verbal symbol learning: A u-shaped trajectory in the acquisition of arbitrary gestures. *Journal of Cognition and Development, 5*, 37–56.
- Opfer, J. E., & Siegler, R. S. (2004). Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology, 49*, 301–332.
- Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology, 28*, 225–273.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford: Oxford University Press.
- Siegler, R. S. (2004). U-shaped interest in U-shaped development—and what it means. *Journal of Cognition and Development, 5*, 1–10.
- Siegler, R. S. (2006). Microgenetic analyses of learning. In W. Damon, & R. M. Lerner (Series Eds.) & D. Kuhn & R. S. Siegler (Eds.), *Handbook of child psychology: Volume 2: Cognition, perception, and language* (6th edn.). Hoboken, NJ: Wiley.
- Siegler, R. S., & Svetina, M. (2002). A microgenetic/cross-sectional study of matrix completion: Comparing short-term and long-term change. *Child Development, 73*, 793–809.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modelling change and event occurrence*. New York: Oxford University Press.
- Taylor, J., & Cox, B. D. (1997). Microgenetic analysis of group based solution of complex two-step mathematical word problems by fourth graders. *Journal of the Learning Sciences, 6*, 183–226.
- Tunteler, E., & Resing, W. C. M. (2002). Spontaneous analogical transfer in 4-year-olds: A microgenetic study. *Journal of Experimental Child Psychology, 83*, 149–166.
- van Geert, P. (2002). Developmental dynamics, intentional action, and fuzzy sets. In N. Granott, & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 319–343). Cambridge: Cambridge University Press.