A New Method for Analyzing Sequential Processes: Dynamic Multilevel Analysis
Ming Ming Chiu and Lawrence Khoo
Small Group Research 2005; 36: 600
DOI: 10.1177/1046496405279309

The online version of this article can be found at:
http://sgr.sagepub.com/cgi/content/abstract/36/5/600

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Small Group Research can be found at:
Email Alerts: http://sgr.sagepub.com/cgi/alerts
Subscriptions: http://sgr.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations http://sgr.sagepub.com/cgi/content/refs/36/5/600
A NEW METHOD FOR ANALYZING SEQUENTIAL PROCESSES: Dynamic Multilevel Analysis

MING MING CHIU
Chinese University of Hong Kong

LAWRENCE KHOO
City University of Hong Kong

Researchers studying sequential processes (e.g., marital conflicts, teacher-student interactions, etc.) often try to model how recent events affect current events. A researcher doing so faces several difficulties: the threat of combinatorial explosion due to comprehensive coding, continuous and discrete variables, and differences across time (nonstationarity) and across groups (group heterogeneity). The authors discuss three often-used methods of analyzing time-series data (conditional probabilities, sequential analysis, and Logit with lag variables) and the problems inherent in them. The authors then introduce a new method that addresses the above problems: dynamic multilevel analysis. To highlight the similarities and differences between these methods, the authors apply them to data from student group problem-solving sessions in an algebra class. The authors use the various methods to show how likelihood of agreement was affected by other recent speakers' correct ideas, mathematics status, agreement, and rudeness.

Keywords: social interaction; sequential analysis; time-series analysis

Many phenomena are inherently sequential. When analyzing series of sequential events (or time-series data), researchers often model how events are affected by recent events within a series. Examples include student interactions with other students or with...
teachers (Chiu, 2004; Chiu & Khoo, 2003), marital conflict (Gottman et al., 2003), and preschoolers’ play (Farran & Son-Yarbrough, 2001). Consider a group of people discussing how to solve a problem. Each person’s action might affect later actions by other group members. For example, a person who states a correct idea likely raises the probability that the next person agrees. This example is a special case of the general phenomenon of a series of interactions among two or more entities (people, animals, countries, etc.).

In this article, we consider the general class of phenomena of current events being affected by recent past events (and also by non-time-dependent characteristics). We examine the difficulties involved in modeling time-series data and several methods for doing so. To concretize the methodological issues, we introduce a specific set of hypotheses and a data set. Then we consider three types of difficulties in analyzing time-series data: coding difficulties, lack of independence among the observations, and heterogeneity in the data set. Next, we review past methods used to model time-series data: conditional probabilities, sequential analysis, and Logit. We then introduce a new method that we call dynamic multi-level analysis (DMA, based in part on Chiu, 2000) and discuss how it mitigates these difficulties. To illustrate the differences among these methods, we use all of them on one data set to test several research hypotheses. We then conclude with a comparison of the various methods.

Before we proceed further, we wish to introduce some terminology. The sequential processes we discuss can involve groups, dyads, or individuals. To avoid confusion, we refer to the object under study (group, dyad, or individual) as the sampling unit. A sampling unit can be observed on one or more occasions; we refer to each such occasion as a session. If warranted, we can further divide each session into time periods. During a session, we observe a stream of sequential behavior. Before analysis, coders parse this stream of sequential behavior into discrete behaviors. We refer to each such discrete behavior as a turn. Often, coders further assign each discrete behavior to one of a set of categories, categorizing the nature of each turn.
EXAMPLE OF
TIME-SERIES DATA AND HYPOTHESES

We begin with a specific set of data and hypotheses drawn from Chiu and Khoo (2003) to contextualize the methodological issues involved. Later, we test these hypotheses by applying all of the various methods to this data.

DATA

The data consist of 80 middle school students’ grades and transcribed videotapes. These students worked on an algebra word problem in 20 groups of 4 students each. The sampling unit for this study is the group. Each group was videotaped for one session. The smallest unit of analysis is a speaker’s turn in a group’s conversation. The transcribed videotapes included 3,104 speaker turns of conversation. Distinct time periods might exist within each session, forming an intermediate level of analysis.

Two research assistants who were unaware of the research hypotheses coded each speaker turn for the following: correctness, speaker’s mathematics status, and evaluation of the previous speaker. A speaker’s mathematics status was computed as his or her mathematics grade minus his or her group’s mean mathematics grade. Evaluations included agreement, polite disagreement, impolite disagreement, neutral actions, and an ignoring of the previous speaker. So the analysis included discrete variables—correct, agree, rudely disagree, and ignore—and a continuous variable, math status. We used Cohen’s (1960) kappa to test for interrater reliability.

HYPOTHESES

In an ideal world, a person agrees when the previous speaker states a correct idea and disagrees otherwise. However, other factors such as status, politeness, and recent agreements might also affect agreement. Controlling for correctness, the previous speaker’s past achievements relative to those of other group mem-
bers (achievement status) may also affect others’ evaluations of his or her idea. Thus, past mathematics grades might bias agreement when a group works on a mathematics problem. Also, rude actions such as ignoring the previous speaker and disagreeing rudely might cause other group members to become defensive and hence less likely to agree with the rude person. Finally, agreement in recent turns may also predict future agreement because previous agreements are likely to build a common knowledge base for agreement in the following turns.

We predict the outcome variable of agree with both turn-level variables and individual-level variables. All turn-level explanatory variables occur before the turn of the outcome variable. Therefore, time constrains the direction of causality. Restating these hypotheses in terms of the variables, we have the following (time lags are in parentheses; \(-1\) indicates the previous turn):

- Previous speaker’s correctness predicts agreement: Correct \((-1 \ldots -4; \text{up to four turns ago})\), correctness of the previous speaker, positively predicts agree \((0)\), agreement by the current speaker.
- Previous speaker’s achievement status predicts agreement: Math status \((-1 \ldots -4)\) positively predicts agree \((0)\).
- Previous speakers’ rude disagreements negatively predict agreement: Rudely disagree \((-1 \ldots -4)\) negatively predicts agree \((0)\).
- Being ignored negatively predicts agreement: Ignore \((-1 \ldots -4)\) negatively predicts agree \((0)\).
- Agreement in recent turns predicts agreement: Agree \((-1 \ldots -4)\) positively predicts agree \((0)\).

**DIFFICULTIES IN ANALYZING TIME-SERIES DATA**

Methods for analyzing time-series data must address at least three types of difficulties: coding difficulties, lack of independence among the observations, and heterogeneity. First, sophisticated hypotheses often require intricate coding of behaviors that threaten consistency and reliability. Second, time-series data observed from different individuals or groups usually violate the independence assumption of many statistics methods. Last, the effect of the
explanatory variables on the outcome variable can differ across groups (group heterogeneity) or change over time (time period heterogeneity or nonstationarity).

CODING DIFFICULTIES

The usefulness of any statistical analysis depends, in part, on the quality of the coding scheme. Ideally, a coding framework for statistical analyses has mutually exclusive and exhaustive categories. Furthermore, these categories should be sufficiently comprehensive to test one’s hypotheses.

However, a complex coding scheme with mutually exclusive and exhaustive categories often includes many codes (e.g., Chiu, 2000). As the number and complexity of the hypotheses rise, the number of codes also rises. This increases the complexity of the coding schemes, the training time for coders, the coding time, and coding conflicts. Thus, complex coding schemes can reduce internal consistency and intercoder reliability.

Moreover, coding schemes with a large number of categories can be statistically problematic. Models with many variables can reduce the available degrees of freedom and the precision of parameter estimates.

NONINDEPENDENT OBSERVATIONS

Researchers studying sequential processes often find that their data violate the assumptions required by traditional statistical models. Methods such as ordinary least squares assume that the model errors are independent between observations (Judge, Griffiths, Hill, Lutkepohl, & Lee, 1985). This assumption is violated by the nature of sequential (or time-series) data, because observations are usually affected by other recent observations. This assumption is also violated when sequential observations are drawn from different sampling units. Observations within a sampling unit can resemble one another substantially and differ from those in other groups in unobserved ways. Ignoring the nonindependence of the observa-
tions can lead to inefficient effect-size estimates and biased estimates of the significance of the explanatory variables.

GROUP HETEROGENEITY AND NONSTATIONARITY

Traditional models also assume that explanatory variable effects are stable over the entire data set. However, this assumption is often violated, because explanatory variable effects can differ across sampling units (also known as group heterogeneity; Goodman, Ravlin, & Schminke, 1987) and also across time (nonstationarity); Dabbs & Ruback, 1987; Goodman et al., 1987). Consider data with multiple sampling units, specifically pairs of students. The interactions between one pair of students can differ substantially from those of other pairs of students. To model these interactions accurately, we need different parameter estimates for each sampling unit.

The effect of an explanatory variable might also change over time. For example, people may agree more often at the end of a problem-solving session than at the beginning of one. They might disagree regularly until they find a correct method (a critical “break point” in the session) and then generally agree afterwards. To model these changes accurately, we need different parameter estimates for each time period.

Also, we might not know the start and end of different time periods (break points). Then, we must identify the break points that divide the sessions into distinct time periods. Specifically, we must estimate the number and locations of break points in the time-series data.

METHODS FOR ANALYZING TIME-SERIES DATA

Researchers have used several methods for analyzing time-series data, including (a) conditional probabilities, (b) sequential analysis, and (c) Logit. In this section, we discuss and apply each method. We then introduce and apply our new method, DMA. Last, we compare these four methods.
Many researchers have used conditional probabilities (CPs) to analyze time-series data, in part because of its simplicity and ease of use (e.g., Farran & Son-Yarbrough, 2001; Parks & Fals-Stewart, 2004; Woods, Rapp, & Beck, 2004). A CP is the probability of an event given that another event has occurred. Using Bayes’s theorem, we can compute them from their overall probabilities (or unconditional probabilities [UPs]). For example, the CP that after a person states a correct idea (C) the next speaker agrees (A) is the quotient of the UP of the sequence CA divided by the UP of A, namely, UP(CA) / UP(A). A table of these CPs can list the likelihoods of sequences of any length occurring after any sequence of earlier events has occurred (see Table 1). The results suggest that correct (–1), agree (–1), and agree (–2) predict agreement.

Using CPs requires little knowledge of statistics, but it has several disadvantages, including the assumption of stationarity, lack of

### TABLE 1: Conditional Probabilities for Antecedents of Agree (0), Whose Unconditional Probability is 56%

| Antecedents of Agree (X) | Unconditional Probability of X | Unconditional Probability of Sequence (X followed by Agree) | Conditional Probability of (X Agree | X) |
|--------------------------|--------------------------------|-------------------------------------------------------------|---------------------------------------|
| Correct (–1)             | 37                             | 24                                                          | 65                                    |
| Agree (–1)               | 56                             | 35                                                          | 64                                    |
| Agree (–2)               | 55                             | 33                                                          | 59                                    |
| Agree (–3)               | 55                             | 34                                                          | 60                                    |
| Agree (–4)               | 55                             | 32                                                          | 59                                    |
| Correct (–1) * Agree (–1)| 21                             | 15                                                          | 71                                    |
| ~Correct (–1) * Agree (–1)| 34                             | 21                                                          | 60                                    |
| Correct (–1) * ~Agree (–1)| 16                             | 9                                                           | 56                                    |
| Correct (–1) * Agree (–2)| 23                             | 15                                                          | 68                                    |
| Correct (–1) * ~Agree (–2)| 14                             | 9                                                           | 61                                    |
| ~Correct (–1) * Agree (–2)| 33                             | 18                                                          | 54                                    |
| Agree (–1) * Agree (–2)  | 35                             | 22                                                          | 63                                    |

...  

NOTE: All values are percentages. We omitted the remainder of this table due to space limitations.
significance tests, discrete variable requirement, and combinatorial explosion. Without a method to identify time periods, a researcher using CPs assumes stationarity. As a crude test, a person using CPs might arbitrarily create time periods and test if the CPs are similar in each time period. Without significance tests, researchers must rely on subjective, human judgment to decide if CPs differ significantly. Also, CPs cannot be computed for continuous variables. Thus, one must accept a loss of precision by dividing a continuous variable into several ranges to create discrete variables. More variables increase the precision but also increase the complexity of the results.

Last, CP results can explode combinatorially due to continuous variables, extra explanatory variables, testing of different sampling units, or testing of time periods. Table 1 shows the potential complexity of the results. For example, a researcher must compute many CPs to determine if an explanatory variable such as correct (−1) has a substantial independent effect or whether its effect stems from a correlation with another explanatory variable. Creating multiple discrete variables from a continuous variable exacerbates this problem. Last, testing for heterogeneity of sampling units or nonstationarity sharply raises the complexity of the results because they require replication of the full set of computations for each sampling unit or each time period.

In short, CP might be sufficient for exploratory analysis. However, its shortcomings render it inadequate for rigorous testing of nontrivial models.

SEQUENTIAL ANALYSIS

Sequential analysis (SA; described in detail in Gottman & Roy, 1990) can be viewed as an advanced version of CP, often used by researchers in various fields (e.g., Han, 2004; Koester, 2004; Lavelli & Fogel, 2005). As with CPs, SA assumes that current actions are probabilistically determined by recent actions. Specifically, SA views sequential phenomena as a discrete Markov process that can take on any one of a finite number of predefined states. The current state determines the probability of the phenomenon
being in a given state in the next period (Papoulis, 1984). SA estimates the transition probabilities between the predefined states. SA also aims to explain asymmetries in these probabilities with a model of explanatory variables.

Unlike CP, SA includes tests of significant differences, effect sizes, and log-linear explanatory models. To test the significance of differences in probabilities, SA compares how much the CP of the consequent behavior, given the antecedent behavior, deviates from the simple UP of the consequent behavior by using the \( z \) score (Allison & Liker, 1982; Bakeman & Gottman, 1986). Bakeman and Quera (1995) showed that the \( z \) score is similar to an adjusted cell residual from a two-way contingency table of the relationship between each behavior and its immediate antecedent. Probit or Logit betas provide an estimate of the effect size of each antecedent (Gottman & Roy, 1990). Last, one can test whether adding an explanatory variable improves the fit of the model by using likelihood ratio chi-square tests (LR\( \chi^2 \); Anderson & Goodman, 1957).

Consider the following application of SA to our data and hypotheses. SA proceeds through the following steps: (a) sample size, (b) coding reliability, (c) order, (d) group heterogeneity, (e) nonstationarity, and (f) log-linear explanatory models.

**Sample Size and Coding Reliability**

For \( s \) states and \( k \) possible lags (or a \( k \) order Markov chain), there are \( s^{k+1} \) cells. The sample size of each subject should exceed 5 times the number of cells, \( 5 \times s^{k+1} \), and 80% of the cells should have an expected probability greater than 5 (Bakeman & Quera, 1995; Tabachnick & Fidell, 1989). Consider the above example with three variables (correctness, evaluation, and speaker status). Correctness has two possible states, right or wrong. Evaluation has four possible states (agree, polite disagree, rude disagree, and ignore). Mathematics status has 33 possible states, but let us simplify them into two states, high and low. Even with this simplification, we have \( 2 \times 4 \times 2 = 16 \) states and four possible lags, so the sample size should exceed \( 5,242,880 \ (5 \times 16^{(4+1)} = 5 \times 16^5) \), far more than the 3,104 data points in the above study. Thus, four possible lags are
not feasible. If the groups are heterogeneous, the suggested sample size applies to each group. As will be shown below, the groups were heterogeneous. The group with the fewest data had 93 data points. So we must reduce the hypotheses to the following first order Markov chain model with only one lag:

- Correct (−1) predicts agree (0).
- Agree (−1) predicts agree (0).

This Correct/Not Correct × Agree/Not Agree model has 4 states, 16 cells \(4^2\), and a suggested sample size of more than 80 \(5 \times 16\). Still, 5 of the 20 groups did not satisfy the heuristic of 80% of the cells having expected probabilities greater than 5. So our analyses proceeded with only 15 groups (new total turns = 2,577). Cohen’s (1960) kappa showed high coding reliability for agree \((k = .97, z = 50.6, p < .001)\).

Tests of Markov Chain Order, Group Heterogeneity, and Nonstationarity

We tested if the order of the Markov chain of each subject was \(k\) using likelihood ratio chi-square tests \(\text{LR}\chi^2;\) Anderson & Goodman, 1957). The remaining 15 groups each showed a significant link between the current state and earlier states. This result suggests at least a first order \((t = 1)\) Markov chain: mean \(\text{LR}\chi^2 = 40.2\), degrees of freedom \((df) = 9, p < .001\). As shown above, the limited sample size did not allow a second order test to be run with confidence. (All groups failed to meet the minimum sample size of 320 = \(5 \times 4^1\).) Running the second order \(\text{LR}\chi^2\) anyway questions the assumption that a first order Markov chain could fit the data could well, because 7 of the 15 groups showed significant second order \((t = 2)\) links (mean for these 7 groups: \(\text{LR}\chi^2 = 191.6, df = 36, p < .001\)).

Next, we test for group heterogeneity and nonstationarity. The results also showed that the groups’ data were significantly heterogeneous for a first order Markov chain, \(\text{LR}\chi^2 = 313, df = 168, p < .001\). So the groups could not be pooled into one data set. When
we divided each group into two equal time periods (assuming a first order Markov chain), 9 of the 15 groups showed significant nonstationarity. The mean for these 9 groups were as follows: $\chi^2 = 76.4$, $df = 48$, $p < .01$. So these nine groups required division into multiple homogeneous time periods for further SA.

The SA results thus far show that the data require further division into smaller data subsets because of multiple Markov chain orders, group heterogeneity, and nonstationarity. We could combine data from homogeneous groups with homogeneous time periods together. The resultant pools of data might be sufficiently large to fit a second order Markov chain (or higher). However, the nine groups with nonstationarity raise a difficult problem. SA does not provide a means for identifying homogeneous time periods. Without identifiable time periods, the data cannot be easily analyzed.

### Explanatory Model and Results

For comparison purposes, we assumed that the groups and time periods are all homogeneous. Probit betas (Gottman & Roy, 1990) and Allison and Liker’s (1982) $z$ were also computed for each model effect. In this case, the possible models predicting agree in the subsequent turn are as follows: (a) only agree ($-1$) (agreement in the prior turn); (b) only correct ($-1$); (c) agree ($-1$) and correct ($-1$); and (d) Agree ($-1$), Correct ($-1$), and Agree ($-1$) $\times$ Correct ($-1$) (interaction term). The following results model 15 groups’ data with a first order Markov chain (see Table 2).

### Table 2: Likelihood Ratio Chi-Square Tests for Four Models for First Order Interactions

<table>
<thead>
<tr>
<th>Likelihood ratio chi-square test</th>
<th>Model 1: Agree ($-1$)</th>
<th>Model 2: Correct ($-1$)</th>
<th>Model 3: Agree ($-1$) $+$ Correct ($-1$)</th>
<th>Model 4: Agree ($-1$) $+$ Correct ($-1$) $\times$ Correct ($-1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degrees of freedom</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>984</td>
<td>773</td>
<td>748</td>
<td>462</td>
</tr>
</tbody>
</table>


Correct (–1) and agree (–1) in the previous turns both predict agreement. Model 4 with the additional interaction term Correct (–1) × Agree (–1) significantly fits the data better than Model 3, which lacks that term (LR $\chi^2$ difference = 286, df difference = 1, $p < .001$). Thus, Model 4 is the best of the four models. Table 3 shows parameter estimates and $z$ values.

SA still suffers from many of the same shortcomings as CP. Combinatorial explosion remains a threat. SA also models a continuous variable by splitting it into one or more discrete variables. Last, SA does not model nonstationarity or sampling unit heterogeneity except through parallel SAs of subsamples of the data.

**REGRESSION ANALYSIS**

Regression analysis takes a different approach to modeling complex phenomena. Unlike CP or SA, regressions can exploit multidimensional coding to create simpler models that capture complex phenomena. After briefly discussing coding schemes, we consider two regression methods, Logit and DMA.

**Multidimensional Coding of Sequential Process Variables**

A multidimensional coding scheme can capture the data’s complexity, reduce the number of needed variables, and increase intercoder reliability. For example, Chiu's (2000) individual action

<table>
<thead>
<tr>
<th>Variable</th>
<th>p(B)</th>
<th>Variable</th>
<th>p(A – 1)</th>
<th>$B \mid A$ (–1)</th>
<th>$B \mid \neg A$ (–1)</th>
<th>Beta</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agree 55%</td>
<td>Correct</td>
<td>36%</td>
<td>58%</td>
<td>42%</td>
<td>0.42</td>
<td>68.9***</td>
<td></td>
</tr>
<tr>
<td>Agree 55%</td>
<td>Agree</td>
<td>55%</td>
<td>63%</td>
<td>37%</td>
<td>0.74</td>
<td>76.3***</td>
<td></td>
</tr>
<tr>
<td>Agree 55%</td>
<td>Correct × Agree</td>
<td>21%</td>
<td>74%</td>
<td>26%</td>
<td>1.37</td>
<td>53.3***</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: A[-1] indicates the variable A at lag 1, namely the likelihood of A by the previous speaker. For example, Correct (-1) refers to the likelihood that the previous speaker was correct.

***$p < .001$. 

Downloaded from [http://sgr.sagepub.com](http://sgr.sagepub.com) at CHINESE UNIV HONG KONG LIB on June 19, 2009
framework has three dimensions. Because each dimension has three categories, this framework can capture 27 different types of action.

By coding one dimension at a time, a coder chooses among only three possible codes (instead of 27). Thus, training and coding time is shorter. Intercoder reliability is also likely to improve. Furthermore, by coding along different dimensions, we do not need state variables. The smaller number of variables facilitates the use of linear explanatory models and methods similar to ordinary least squares (OLS). In short, using multiple dimensions retains the complexity of the categories, likely improves internal consistency and intercoder reliability, and facilitates the use of linear models.

**Logit**

Many sequential analyses have discrete outcome variables (e.g., agree or not agree). When the outcome variable is discrete instead of continuous, simple regression methods such as OLS are not suitable, because OLS is inefficient and yields biased results. Researchers have used Logit instead to analyze time-series data with discrete outcome variables (Gupta, 1988; Pevalin & Ermisch, 2004; Silverstein & Parker, 2002). Logit is used to estimate how the probability of observing an event (e.g., agree or not agree) is related to various continuous or discrete explanatory variables (for details, see Greene, 1997). In our example, the probability of current agreement might depend on mathematics status (−1), correct (−1), or agree (−1).

A Logit model assumes that there is a continuous, unobserved, outcome variable, \( y' \) that is linearly related to various explanatory variables. The unobserved variable, \( y' \), is in turn linked to the observed discrete outcome variable, \( y \), by a link function. This link function describes the probability that the observed outcome variable takes on a particular value, given the value of the unobserved variable. A simple, often-used link function is the Logit link function. To ensure that the results do not depend on the particular distribution of the Logit link function, researchers often also use other
link functions with different distributions (e.g., Probit, Gompit, etc.) as tests of robustness.

*Logit description.* The unobserved variable in the Logit model can be described as

\[ y_i^* = \beta_0 + \beta_1 x_i + e_i \]  

(1)

To model lag effects, we can enter lag variables as explanatory variables, such as correct (–1).

\[ y_i^* = \beta_0 + \beta_1 x_i \text{**Correct}_{i-1} + e_i \]  

(2)

Let \( \pi_{it} \) be the probability that an event (e.g., agreement) is observed for at time \( t \). It is determined by the expected value of the unobserved variable and the Logit link function:

\[ \Pi_t = p(y_i = 1|x_i; \beta_0, \beta_1) = F(\beta_0 + \beta_1 x_i) \]  

(3)

\[ \frac{1}{1 + e^{-\beta_1 x_i}} \]  

(4)

After estimating the regression results, we facilitate their interpretation by converting the coefficients into percentage change in the probability of the outcome variable. For Logit,

\[ \ln \left( \frac{p}{1-p} \right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \]  

(5)

We take the antilogs and solve for \( p_1 - p_0 \) (\( p_1 \) and \( p_0 \) are the probabilities of the outcome variable when \( x_i = 1 \) and \( 0 \), respectively) to obtain

\[ p_1 - p_0 = \frac{1}{1 + e^{(b_0 + b_1 x_1 + \ldots + b_n x_n)}} - \frac{1}{1 + e^{(b_0 + b_1 x_1 + \ldots + b_n x_n)}} \]  

(6)
Likewise for a continuous variable, the percentage change in the probability of the outcome variable for each unit of the explanatory variable is

\[
\frac{dp}{dx_i} = \frac{b_i e^{(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_n x_{in})}}{1 + e^{(b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_n x_{in})}}
\]

(7)

For simplicity, we discuss only binary outcome variable regressions. See Greene (1997) for a discussion of multinomial and ordered outcome variables.

Applying Logit. To do an analysis, we check the sample size requirement, test the intercoder reliability, specify the explanatory variables for the Logit regression, and repeat the analyses with Probit and Gompit. The sample size requirement for regressions is much smaller than that for SA. Green (1991) proposed the following heuristic sample size, \(N\), for a multiple regression with \(M\) explanatory variables and an expected effect size of \(R^2\):

\[
N > 8 \left(1 - R^2\right) / R^2 + M - 1
\]

To test our hypotheses, our model requires five variables with up to four lags, or 20 (5 \times 4) variables. So the model requires 811 data points if the expected effect size is very small; for \(R^2 = .01\) and \(M = 20\), \(N = 8(0.99)/0.01 + 20 - 1 = 811\). Hence, the actual data set of 3,104 is sufficient. (If the expected effect size is larger, the sample size can be smaller. If the expected \(R^2 = .05\), the suggested sample size is only 171.)

Intercoder reliability was acceptable. Evaluations included 54% agreements, 0.3% neutral, 16% ignore/unresponsive turns, 18% polite disagreements, and 9% rude disagreements (Cohen’s kappa = .93, \(z = 49.5\), \(p < .001\)).

Next, we add sets of explanatory variables to the model, prioritized by theoretical importance and time. Correct (-1) is likely to have the strongest effect, so we enter it first. Because the more recent explanatory variables likely have the strongest effects, we enter the most recent remaining explanatory variables, at Lag 1,
mathematics status (–1), agree (–1), ignore (–1), and rudely disagree (–1). Next, we add all the explanatory variables at Lag 2, then at Lag 3, and finally at Lag 4. Last, we test for interaction effects among the significant explanatory variables.

Table 4 shows that the following properties of previous speakers positively predicted agreement: correct (–1), mathematics status (–1), and agree (–1). Meanwhile, rudely disagree (–1, –2, –3) negatively predicted agreement (see Table 1). Other explanatory variables and all interaction effects were not significant. The discrete explanatory variables affected the likelihood of the current speaker

### Table 4: Summaries of Five Binary Logit Regression Models Predicting Agreement With Unstandardized Coefficients, Standard Errors, and the Percentage Change in the Probability of the Outcome Variable Agree

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>B</th>
<th>SE</th>
<th>β</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.587</td>
<td>0.076</td>
<td>0.587</td>
<td>14.2***</td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.561</td>
<td>0.077</td>
<td>0.561</td>
<td>13.6***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.015</td>
<td>0.003</td>
<td>0.186</td>
<td>0.8***</td>
</tr>
<tr>
<td><strong>Step 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.518</td>
<td>0.079</td>
<td>0.518</td>
<td>11.6***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.014</td>
<td>0.003</td>
<td>0.174</td>
<td>0.8***</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.665</td>
<td>0.080</td>
<td>0.665</td>
<td>16.5***</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>–0.504</td>
<td>0.139</td>
<td>–0.504</td>
<td>–12.5***</td>
</tr>
<tr>
<td><strong>Step 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.525</td>
<td>0.079</td>
<td>0.525</td>
<td>11.6***</td>
</tr>
<tr>
<td>Math status (–1)</td>
<td>0.013</td>
<td>0.003</td>
<td>0.162</td>
<td>0.7***</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.647</td>
<td>0.081</td>
<td>0.647</td>
<td>16.0***</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>–0.455</td>
<td>0.141</td>
<td>–0.455</td>
<td>–11.2**</td>
</tr>
<tr>
<td>Rudely disagree (–2)</td>
<td>–0.435</td>
<td>0.131</td>
<td>–0.435</td>
<td>–10.7**</td>
</tr>
<tr>
<td><strong>Step 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.519</td>
<td>0.080</td>
<td>0.519</td>
<td>11.3***</td>
</tr>
<tr>
<td>Math status (–1)</td>
<td>0.012</td>
<td>0.003</td>
<td>0.012</td>
<td>0.7***</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.620</td>
<td>0.082</td>
<td>0.620</td>
<td>15.3***</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>–0.437</td>
<td>0.142</td>
<td>–0.437</td>
<td>–10.7**</td>
</tr>
<tr>
<td>Rudely disagree (–2)</td>
<td>–0.368</td>
<td>0.133</td>
<td>–0.368</td>
<td>–9.0**</td>
</tr>
<tr>
<td>Rudely disagree (–3)</td>
<td>–0.593</td>
<td>0.133</td>
<td>–0.593</td>
<td>–14.6***</td>
</tr>
</tbody>
</table>

**NOTE:** Each regression included a constant term. For Step 1, McFadden’s $R^2 = .014$, Akaike information criterion (AIC) = 1.354; for Step 2, $ΔR^2 = .006$, $ΔAIC = –.007$; for Step 3, $ΔR^2 = .028$, $ΔAIC = –.037$; for Step 4, $ΔR^2 = .004$, $ΔAIC = –.005$; for Step 5, $ΔR^2 = .004$, $ΔAIC = –.006$.

* $p < .05$. ** $p < .01$. *** $p < .001$. 

*Downloaded from http://sgr.sagepub.com at CHINESE UNIV HONG KONG LIB on June 19, 2009*
agreeing substantially, ranging from $-14.6\%$ (rudely disagree, $-3$) to $+15.3\%$ (agree, $-1$). Meanwhile, each 1 point increase in the previous speaker’s mathematics status raised the likelihood of agreement by the current speaker by 0.8%.

Rudely disagree showed the strongest and longest lasting effects. The coefficients for rudely disagree ($-1$, $-2$, and $-3$) cumulatively exceeded correct ($-1$). Rudely disagree also affected the likelihood of agree three turns later, whereas the other effects lasted only one turn. The Probit and Gompit analyses yielded similar results, showing that these results did not stem from the distribution assumptions (results are not shown due to space considerations; results are available from authors upon request).

Similar to SA, Logit yields estimates of effect sizes and degree of model fit to the data. Unlike SA, Logit can test complex sets of hypotheses with relatively small data sets, and it can model continuous explanatory variables. Applying Logit to subsamples of the data serves as crude tests of nonstationarity or heterogeneity, but it does not model them. Logit also assumes a linear explanatory model and independent and identically distributed errors.

**DYNAMIC MULTILEVEL ANALYSIS**

Similar to Logit, DMA is based on regression analysis. It proceeds through three steps: (a) identification of time periods via break points, (b) multilevel Logit, and (c) testing for serial correlation and modeling it if necessary. Although the components in this technique are not new, combining them in this fashion addresses shortcomings of the previous methods.

**Estimating Break Points to Identify Time Periods**

Suppose variable effects do not remain stable over time (non-stationary). We can model time period differences by dividing the data into different distinct time periods and applying multilevel analysis. First, we must identify the number and locations of break points that divide a session into different time periods. Maddala and Kim (1998) viewed the estimation of an unknown number of
break points as a model selection problem and argued that the optimal model has the lowest Schwarz Bayesian information criterion (BIC). Conceptually, information criteria measure whether a model strikes a good balance between goodness of fit and a parsimonious specification. The BIC is defined as

\[
-\frac{2L}{n} + \left(\frac{k \ln(n)}{n}\right)
\]

where \(k\) is the number of estimated parameters, \(n\) is the number of observations, and \(L\) is the value of the log likelihood function using the \(k\) estimated parameters. Hence, the BIC is a measure of goodness of fit, adjusted by a penalty that increases with the number of regressors in the model. Many statistics software packages compute BIC for each model, allowing us to select the one with the lowest BIC as the optimal model (for more details, see Grasa, 1989).

We use a modified version of Maddala and Kim’s (1998) method to identify the break points for each session. Using only the outcome variable \(y_t\), we create a simple univariate time-series model (an auto-regressive Order 1 model).

\[
y_t = C + \beta y_{t-1} + \epsilon_t
\]

With break points this model becomes

\[
y_t = C + C_2d_2 + C_3d_3 + \ldots + C_pd_p + \beta y_{t-1} + \epsilon_t
\]

where \(p\) is the number of time periods and \(d_p\) is a dummy variable associated with time period \(p\).

We begin by assuming a maximum possible number of break points (guided by theoretical and computational considerations). Using data from one session, we compute the BIC in Equation 9 with no break points. Next, we use Equation 10 to compute the BICs for all possible locations of one break point in this time series (calculate the BIC if the break is between Observation 1 and Observation 2, then if it is between Observation 2 and Observation 3,
etc.) Then, we compute the BICs for all possible combinations of two break points, all possible combinations of three break points, and so on until we have done so for the maximum number of break points.

The model with the lowest BIC has the optimal number and position(s) of the break point(s). Repeating this procedure for each session identifies the number and locations of the break points in all sessions.

Researchers may have a priori information on break points. If the break points are known, researchers can proceed directly to the multilevel analysis below. If they have specific candidates, they can compare them with the break points estimated by this procedure, choose appropriate break points, and proceed with the multilevel analysis.

Multilevel Analysis

We can model the differences across groups and across time periods by using multilevel analysis (Goldstein, 1995; also known as hierarchical linear modeling, Bryk & Raudenbush, 1992). Specifically, we can use it to address sampling unit-specific, session-specific, and time-period-specific effects. A multilevel model estimates the relationship between an outcome variable and sets of explanatory variables defined at different levels of analysis.

In our data, the sampling unit is the group. Each group is observed for only one session, so session is not a separate level of analysis. A speaker turn occurs within a specific time period, which occurs within a particular group, showing a nested hierarchical structure. Hence, there are three levels of analysis, speaker turn at the lowest level, followed by time period at the next level and group at the highest level.

A basic two-level model. The following illustrates the structure of multilevel models using a simple two-level example—speaker turns collected from different groups, with no time periods. Speaker turn variables are at Level 1. Group variables are at Level 2.
For the moment, let us use only continuous variables. Let the outcome variable be the speaker’s fluency score for that turn \( (y_i) \). Let the explanatory variable be the number of words in that turn \( (x_i) \). The simple regression relationship between outcome and explanatory variables for each speaker turn is

\[ y_i = \beta_0 + \beta_1 x_i + e_i \]  

(11)

The subscript \( i \) takes values from 1 to \( n_j \), where \( n_j \) is the number of speaker turns in a group.

Because there are multiple groups, we allow the relationships to differ among the groups. We can express the multiple relationships as

\[ y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \]  

(12)

The subscript \( j \) takes a different value for each group.

In multilevel analysis, the observed Level 2 groups are viewed as a random sample of all possible groups. The parameters in each group \( (\beta_{0j} \text{ and } \beta_{1j}) \) deviate from the global parameters of all groups \( (\beta_0 \text{ and } \beta_1) \) by the random residuals \( (u_{0j} \text{ and } u_{1j}) \). So, the full multilevel model is

\[ y_{ij} = (\hat{\beta}_0 + u_{0j}) + (\hat{\beta}_1 + u_{1j}) x_{ij} + e_{ij} \]  

(13)

or

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij} + e_{ij} \]  

(14)

This equation can be divided into fixed components \( (\beta_0 \text{ and } \beta_1) \) and random components \( (u_{0j}, u_{1j}, \text{ and } e_{ij}) \). The random parameters all have means of zero. We assume that these random variables are not correlated with one another and that they follow a normal distribution. So, it is sufficient to estimate their variances, \( \sigma^2_{u0}, \sigma^2_{u1}, \text{ and } \sigma^2_{e} \), respectively. We can estimate the parameters of a multilevel
model (such as in the above model, \( \beta_0, \beta_1, \sigma^2_{\epsilon}, \sigma^2_{u}, \sigma^2_{v} \)) using maximum likelihood methods (see Goldstein, 1995).

**A three-level model for heterogeneous groups and time periods.**

If the data differ across groups and across time periods, we need a three-level model. The lowest level is the speaker turn. The second level is the time period, and the third level is the group.

A full three-level, multilevel model with a single explanatory variable becomes

\[
y_{ijk} = (\beta_0 + u_{0jk}) + (\beta_1 + u_{1jk})x_{ijk} + e_{ijk}
\]

or

\[
y_{ijk} = \beta_0 + \beta_1 x_{ijk} + u_{0jk} + u_{1jk}x_{ijk} + v_{0k} + v_{1k}x_{ijk} + e_{ijk}
\]

The subscripts, \( i, j, \) and \( k \), refer to speaker turn \( i \) in the \( j \)th time period from \( k \)th group. The subscript \( i \) takes values from 1 to \( n_{jk} \), where \( n_{jk} \) is the number of speaker turns (Level 1) in the \( j \)th time period of the \( k \)th group. The subscript \( j \) takes values from 1 to \( n_k \), where \( n_k \) is the number of time periods (Level 2) in the \( k \)th group. And last, the subscript \( k \) takes a different value for each group (Level 3).

**Practical considerations.** Modeling data from different time periods, groups, and/or sessions does not always require multilevel analysis (Goldstein, 1995). The time periods, groups, or sessions could be homogenous. We can test for heterogeneity by running a model with only a constant term and random terms for each level of analysis (time period \( u_{0jk} \), group \( v_{0k} \), and session \( w_{0l} \); i.e., a variance components model):

\[
y_{ijkl} = \beta_0 + u_{0ijkl} + w_{0lijkl} + e_{ijkl}
\]

If none of the random terms are statistically significant, there is no evidence for significant time period, group, or session differences. Then, a multilevel model is not needed and OLS or Logit would
suffice. If any of the random terms are significant, a multilevel analysis is needed to adequately model the data.

The power issues are complicated, but the following can serve as rules of thumb (Goldstein, 1995). The sample size at the highest level (in this case, the group level) should be as large as possible, at least 20. Moreover, each lower level should average at least 5 times more data points than the level above it (although ratios as small as 3 have been successful; Chiu & Khoo, 2003). At the lowest level of a multilevel analysis, however, the number of data points for a specific group or time period can be very small, even just one (Braun, Jones, Rubin, & Thayer, 1983).

**Logit and multilevel models.** For discrete outcome variables, we combine a Logit type model with multilevel analysis. The resulting multilevel Logit models can be estimated using quasi-maximum likelihood techniques (see Goldstein & Rasbash, 1996). Conceptually, a multilevel Logit model can be divided into its multilevel part and its Logit part. Consider a basic two-level model with an unobserved variable, \( y_{ij}^* \) at turn \( i \) of group \( j \), a single explanatory variable, \( x_{ij} \), and a Logit link function:

\[
y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}
\]  

(18)

The probability (\( \pi_{ij} \)) that an event (e.g., agreement) occurs at turn \( i \) of group \( j \) is determined by the expected value of the unobserved variable and the Logit link function:

\[
\pi_{ij} = p(y_{ij} = 1 \mid x_{ij}, \beta_0, \beta_1 = F(\beta_0 + \beta_1 x_{ij} + \mu_j))
\]  

(19)

\[
\pi_{ij} = \frac{1}{1 + e^{-\beta_0 - \beta_1 x_{ij} - u_j}}
\]  

(20)

The Level 2 variation parameter \( u_j \) represents the deviation of group \( j \) from the population norm. The Level 1 variation, \( e_{ij} \), does not contribute to the fixed components and is a random variable.
only at Level 1. So we constrain the variance of $e_{ij}$ to 1 without loss of generality.

Therefore, the observed outcome variable $y_{ij}$ is

$$y_{ij} = \pi_{ij} + e_{ij} + z_{ij}$$  \hspace{1cm} (21)

$$\sigma^2_e = 1$$  \hspace{1cm} (22)

$$z_{ij} = \sqrt{\pi_{ij}(1-\pi_{ij})}$$  \hspace{1cm} (23)

We can estimate the above parameters ($\beta_0$, $\beta_1$, and $\sigma^2_e$) with either predictive or marginal quasi-likelihood methods (see Goldstein & Rasbash, 1996). By repeating the analysis with multilevel Probit or Gompit, we can check if the results rely on Logit distribution assumptions.

**Serial Correlation**

In time-series data, events might resemble other recent events (serial correlation). For example, agreement during group problem solving tends to occur in clumps. Conversations might move between topics in which speakers mostly agree and topics whereby they mostly disagree. If the time-series relationships are not modeled properly, serial correlation in the data can affect model parameter estimates. When model residuals are serially correlated, the model is likely inappropriate due to inefficient parameter estimates and biased estimates of the parameters’ standard errors (Enders, 1995; Hamilton, 1994). For example, OLS is a minimum variance unbiased estimator only if its residuals are independent and identically distributed (also known as white noise). The residual terms’ values must not be serially correlated—that is, correlated with its earlier or later values.

Our model tries to capture all time-series effects, for example, by including lagged variables as explanatory variables if appropriate. A model that includes lags of both the explanatory variables ($x_i$) and the outcomes variable ($y_{i-1}$) has the following form:
If done successfully, the model's residuals are white noise. We use a method based on the BIC to estimate the appropriate time-series model and to choose the appropriate number of lagged variables. Unlike other information criteria, the BIC provides a consistent estimator for the number of lagged variables in the true model (Grasa, 1989; Lütkepohl, 1985). We proceed as follows (Enders, 1995):

1. Regress the outcome variable on all reasonable lags of the explanatory variables (and possibly lagged terms of the outcome variable).
2. Choose the model that minimizes the BIC (without dropping terms from the middle of a lag structure).
3. Test for serial correlation in the residuals and accept the model if residuals are white noise.

We use Ljung-Box $Q$ statistic (Ljung & Box, 1979) to test for serial correlation in the residuals of the regression with the lowest BIC. The $Q$ statistic at lag $k$ tests the null hypothesis that there is no autocorrelation up to order $k$. The $Q$ statistic is defined as follows:

$$Q = T(T + 2) \sum_{j=1}^{k} \frac{r_j^2}{T-j}$$

where $T$ is the number of observations and $r_j$ is the $j$th autocorrelation. The $j$th autocorrelation, $r_j$, is estimated as

$$r_j = \frac{\sum_{t=k+1}^{T} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2}$$

If the residuals are serially correlated, the time-series model must be modified, usually by adding more explanatory variables and/or
longer lags of the explanatory variables already in the model. The $Q$ statistic for the final model should not reject the hypothesis that the residuals are not serially correlated. If the residuals are not serially correlated, the time-series model is likely suitable and estimates are likely unbiased. This also applies to Logit and Probit models (Robinson, 1982).

Applying DMA

The analysis proceeded in the order described above. After computing the minimum sample size and intercoder reliability as with Logit, we identified the time periods and break points. Then, we tested for heterogeneity across groups and across time periods with a three-level Logit variance components model using the software MLn (Rasbash & Woodhouse, 1995). After the result showed significant variance at all three levels, we added explanatory variables to a three-level Logit model in the same manner as the simple Logit model but allowing for coefficients to differ across groups and across time periods (random effects). Third, we tested for residual serial correlation with $Q$ statistics. If we found serial correlation, we would have modified the model to eliminate the serial correlation. In this specific study, there was no serial correlation in the proposed model.

As with Logit, we facilitated interpretation of the results by converting the coefficients into percentage of change in the probability of the outcome variable. To ensure that the results did not depend on the Logit distribution assumptions, the entire analysis was repeated with multilevel Probit. (Due to space considerations, we did not include these results, but they are available upon request.)

As noted above in the Logit regression, the sample size and intercoder reliability were adequate. The estimation of break points yielded one to four break points for each group, resulting in two to five different time periods. Thus, the tendency to agree differed across time periods within each session.

The variance components model showed significant variation of agree at both the group level ($0.14, SE = 0.05$) and the time period level ($0.06, SE = 0.02$). So the groups and the time periods were
both heterogeneous with respect to agree. Thus, a three-level multilevel model with groups, time periods, and speaker turns was suitable. Of the total agree variance, 12% occurred at the group level, and 5% occurred at the time period level.

A three-level multilevel time-series Logit model with the lowest BIC was selected. Table 5 shows that the following properties of previous speakers positively predicted agreement (lags in paren-

### Table 5: Summaries of Five Multilevel Logit Regression Models Predicting Agreement With Unstandardized Coefficients, Standard Errors, and the Percentage Change in the Probability of the Outcome Variable Agree

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>B</th>
<th>SE</th>
<th>β</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.445</td>
<td>-0.084</td>
<td>0.445</td>
<td>+10.6***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.011</td>
<td>-0.005</td>
<td>0.137</td>
<td>+0.6*</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.438</td>
<td>-0.084</td>
<td>0.438</td>
<td>+10.5***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.010</td>
<td>-0.005</td>
<td>0.124</td>
<td>+0.6*</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.333</td>
<td>-0.105</td>
<td>0.333</td>
<td>+8.3**</td>
</tr>
<tr>
<td>Ignore (–1)</td>
<td>-0.313</td>
<td>-0.132</td>
<td>-0.313</td>
<td>-7.8*</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>-0.819</td>
<td>-0.160</td>
<td>-0.819</td>
<td>-20.2***</td>
</tr>
<tr>
<td>Step 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.426</td>
<td>-0.086</td>
<td>0.426</td>
<td>+9.9***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.010</td>
<td>-0.005</td>
<td>0.124</td>
<td>+0.6*</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.333</td>
<td>-0.105</td>
<td>0.333</td>
<td>+8.3**</td>
</tr>
<tr>
<td>Ignore (–1)</td>
<td>-0.313</td>
<td>-0.132</td>
<td>-0.313</td>
<td>-7.8*</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>-0.779</td>
<td>-0.161</td>
<td>-0.779</td>
<td>-19.2***</td>
</tr>
<tr>
<td>Rudely disagree (–2)</td>
<td>-0.411</td>
<td>-0.136</td>
<td>-0.411</td>
<td>-10.2**</td>
</tr>
<tr>
<td>Step 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.429</td>
<td>-0.086</td>
<td>0.429</td>
<td>+9.9***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.009</td>
<td>-0.005</td>
<td>0.112</td>
<td>+0.5*</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.313</td>
<td>-0.105</td>
<td>0.313</td>
<td>+7.7**</td>
</tr>
<tr>
<td>Ignore (–1)</td>
<td>-0.315</td>
<td>-0.132</td>
<td>-0.315</td>
<td>-7.8*</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>-0.779</td>
<td>-0.161</td>
<td>-0.779</td>
<td>-19.2***</td>
</tr>
<tr>
<td>Rudely disagree (–2)</td>
<td>-0.411</td>
<td>-0.136</td>
<td>-0.411</td>
<td>-10.2**</td>
</tr>
<tr>
<td>Rudely disagree (–3)</td>
<td>-0.345</td>
<td>-0.138</td>
<td>-0.345</td>
<td>-8.5*</td>
</tr>
<tr>
<td>Step 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct (–1)</td>
<td>0.429</td>
<td>-0.086</td>
<td>0.429</td>
<td>+9.8***</td>
</tr>
<tr>
<td>Math Status (–1)</td>
<td>0.009</td>
<td>-0.005</td>
<td>0.112</td>
<td>+0.5*</td>
</tr>
<tr>
<td>Agree (–1)</td>
<td>0.306</td>
<td>-0.105</td>
<td>0.306</td>
<td>+7.5**</td>
</tr>
<tr>
<td>Ignore (–1)</td>
<td>-0.304</td>
<td>-0.132</td>
<td>-0.304</td>
<td>-7.5*</td>
</tr>
<tr>
<td>Rudely disagree (–1)</td>
<td>-0.757</td>
<td>-0.161</td>
<td>-0.757</td>
<td>-18.7***</td>
</tr>
<tr>
<td>Rudely disagree (–2)</td>
<td>-0.345</td>
<td>-0.138</td>
<td>-0.345</td>
<td>-8.5*</td>
</tr>
<tr>
<td>Rudely disagree (–3)</td>
<td>-0.568</td>
<td>-0.138</td>
<td>-0.568</td>
<td>-14.1***</td>
</tr>
</tbody>
</table>

NOTE: Each regression included a constant term. For Step 1, log likelihood \( (L) = 4051 \), explained group variance \( (V_g) = 33\% \), explained time period variance \( (V_t) = 4\% \); for Step 2, \( \Delta L = -17 \), \( \Delta V_g = 2\% \), \( \Delta V_t = 3\% \); for Step 3, \( \Delta L = -80 \), \( \Delta V_g = 16\% \), \( \Delta V_t = 32\% \); for Step 4, \( \Delta L = -12 \), \( \Delta V_g = 2\% \), \( \Delta V_t = 4\% \); for Step 5, \( \Delta L = -22 \), \( \Delta V_g = 2\% \), \( \Delta V_t = 6\% \). *p < .05. **p < .01. ***p < .001.
theses): correct (–1), math status (–1), and agree (–1). Meanwhile, ignore (–1) and rudely disagree (–1, –2, –3) negatively predicted agreement. All other predictors and all interaction effects were not significant. As with Logit, the discrete explanatory variables affected the likelihood of the current speaker agreeing substantially, ranging from –18.7% (rudely disagree, –1) to +9.8% (correct, –1). Also, each 1 point increase in the previous speaker’s mathematics status raised the likelihood of agreement by the current speaker by 0.5%.

Rudely disagree showed the strongest and longest lasting effects. The coefficients for rudely disagree (–1 and –3) were the largest, exceeding even correct (–1). Rudely disagree also affected the likelihood of agree three turns later, whereas the other effects lasted only one turn.

None of the variances of the explanatory variables’ coefficients were significant at either the group or time period level. This result suggests that the explanatory variables’ effects on agreement are general across both groups and time periods. After testing nested hypotheses of successive deletions of nonsignificant explanatory variables (with Wald tests; see Davidson & MacKinnon, 1993), the same significant explanatory variables remain. The Q statistics of the final model showed no significant serial correlation of residuals in any of the 20 groups. So the time-series model is likely appropriate. Repeating the analysis with multilevel Probit showed similar results.

The results show that as expected, students are more likely to agree with a correct idea than an incorrect one. However, mathematical status, recent agreement, and rudeness all distorted their evaluations of one another’s ideas. The effect of rudeness on agreement was larger and longer lasting than that of correctness.

COMPARISON OF THE DIFFERENT METHODS

Comparison of Results

The results from using the four methods showed some consistencies but several inconsistencies. They all showed that both cor-
rectness and agreement by the previous speaker raised the likelihood of agreement by the current speaker. Because of the sheer number of conditional probabilities, comparing the effects of multiple explanatory variables was difficult. Meanwhile, SA assumptions were not valid for more complex models because they required substantially more data. Also, the Correct (–1) × Agree (–1) interaction term was significant in SA but not in Logit or DMA. The Logit results resembled the DMA results with some exceptions. First, Logit did not find a significant effect for ignore (–1), a result that was evident from DMA. Also, the effect sizes differed substantially. For example, Logit underestimated the effects of rudely disagree, as found in DMA.

CP, SA, and Logit do not properly test for nonstationarity. As suggested by the results from the break point estimation, a LR$\chi^2$ test on two equal time periods does not adequately test for nonstationarity. According to the SA’s LR$\chi^2$ test results, only 6 of the 15 groups had stationarity. However, the break point method divided each group into at least two different time periods, showing that all groups had heterogeneous time periods.

**Suitability of Each Method**

Retaining some basic regression assumptions, DMA addresses several major difficulties involved in modeling the effects of recent events on subsequent events within a series (see comparison Table 6). Similar to other regressions analyses, DMA assumes a linear explanatory model and independent and identically distributed errors. With the proper theoretical framework and coding, however, DMA addresses the following difficulties: (a) the threat of combinatorial explosion, (b) modeling both continuous and discrete variables, and (c) modeling differences across time (nonstationarity) and across groups (group heterogeneity).

When testing complex models, combinatorial explosion of possible states threatens CP and SA but not simple regressions or DMA. As a result, SA can require prohibitively large sample sizes to yield suitable estimates of effect size and model fit. (Because CP does not produce these estimates, it does not require large sample
sizes.) In contrast, proper coding can yield relatively few explanatory variables and modest sample size requirements for simple regressions and DMA. (Simple regressions and DMA also produce estimates of effect size and model fit.)

CP and SA substitute discrete variables for continuous variables, causing loss of precision. In contrast, simple regressions and DMA allow both discrete and continuous variables as both explanatory and outcome variables.

Unlike the other methods, DMA can identify time periods, model heterogeneous time periods and sampling units, and test for serial correlation. DMA can also model serial correlation, if necessary. CP and simple regressions (with the help of statistics software) are relatively easy to use. In contrast, SA and DMA are substantially more difficult to use.

**CONCLUSION**

Past studies have used conditional probability, sequential analysis, and Logit models when studying dynamic processes in which

---

**TABLE 6: A Comparison of Conditional Probabilities (CP), Sequential Analysis (SA), Simple Regressions and Dynamic Multilevel Analysis (DMA)**

<table>
<thead>
<tr>
<th>Properties</th>
<th>CP</th>
<th>SA</th>
<th>Simple Regression</th>
<th>DMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model complex interactions</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Small sample size</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Estimate effect sizes and model fit</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Discrete outcome variables</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Discrete explanatory variables</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Continuous outcome variables</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Continuous explanatory variables</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model nonstationarity</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model group heterogeneity</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Identify different time periods</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Model serial correlation of errors</td>
<td>NA</td>
<td>NA</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td>Easy to use</td>
<td>✓</td>
<td>X</td>
<td>✓</td>
<td>X</td>
</tr>
</tbody>
</table>

_NOTE: OLS = ordinary least squares._

Downloaded from http://sgr.sagepub.com at CHINESE UNIV HONG KONG LIB on June 19, 2009
events in a series are affected by recent past events. Coding difficulties and modeling of only discrete explanatory variables limit the applicability of the conditional probability and sequential analysis methods, enabling them to model only simple phenomena. Furthermore, all three methods fail to address heterogeneity across time or across groups.

We introduced and applied a method that addresses all of these problems: DMA. Given a small data set, multidimensional coding allows models with more explanatory variables, especially those with longer lags (from earlier turns). DMA allows both discrete and continuous variables as both outcome and explanatory variables. Using break point estimation and multilevel techniques, DMA also models the heterogeneity across time periods, groups, and sessions.

NOTE

1. The initial three-level variance components model tests for significant variance at the group, time period, and turn levels:

\[ y_{ijk} = \beta_0 + u_{0jk} + v_{0k} + e_{ijk} \]

\[ \pi_{ijk} = p(y_{ijk} = 1|\beta_0, u_{jk}, v_k = F(\beta_0 + u_{jk} + v_k)) \]

\[ = \frac{1}{1 + e^{-\beta_0 + u_{jk} + v_k}} \]

REFERENCES


Ming Ming Chiu (Ph.D., University of California, Berkeley, 1996) is an associate professor of educational psychology at the Chinese University of Hong Kong. He examines classroom interactions and designs statistical analyses of complex data. Recently, he published several articles on students' interactions during group problem solving and on the academic achievement of 193,076 students from 41 countries.

Lawrence Khoo (Ph.D., Harvard, 1996) is a lecturer of economics at City University of Hong Kong. He does research on statistical methods, fertility, and resource distribution. Recently, he examined the effects of status and politeness on group problem solving, developed an overlapping-generations model for changes in women's fertility, and examined cross-country differences among students using large-scale international studies.