Running Head: HIERARCHICAL MODELING

Hierarchical Modeling of Sequential Behavioral Data: An Empirical Bayesian Approach *

Getachew A. Dagne
University of South Florida

George W. Howe
The George Washington University

C. Hendricks Brown
University of South Florida

Bengt O. Muthén
University of California, Los Angeles

Abstract

This paper reviews the common methods for measuring strength of contingency between two behaviors in a behavioral sequence, the binomial z-score and the adjusted cell residual, and points out a number of limitations with these approaches. It presents a new approach using log odds ratios and employing empirical Bayes estimation in the context of hierarchical modeling, an approach not constrained by these limitations. A series of hierarchical models is presented to test the stationarity of behavioral sequences, the homogeneity of sequences across a sample of episodes, and whether covariates can account for variation in sequences across the sample. These models are applied to observational data taken from a study of the behavioral interactions of 254 couples, to illustrate their use.

Introduction

Behavior is inherently sequential. It unfolds over time, one action following another in a constant stream. When two or more people are together, their individual streams of behavior can intertwine, forming broad rivers of interaction. For several decades behavioral researchers have struggled to characterize these streams of individual behavior or group interaction, developing observational methods for parsing this flow into meaningful units, and constructing quantitative indicators to capture and compare patterns within this stream. These techniques are now used to study questions such as how parent-infant interaction affects attachment security (Kiser, Bates, Maslin, & Bayles, 1986), how parent-child interaction shapes the development of aggressive behavior (Patterson, 1979), how peer-peer interaction influences risky behavior (Bank, Patterson, & Reid, 1996), and how husband-wife interaction influences satisfaction with the relationship (Gottman, 1979).

In this paper we are concerned with quantitative methods used to describe recurrent regularities in microcoded observational data. In microcoding, the stream of behavior or interaction is first parsed into discrete behaviors, and then each behavior is assigned to one of a set of exhaustive categories.

We begin this paper by describing quantitative methods that have been developed to characterize patterns within such data sets. We then discuss important limitations of these methods, including their over-sensitivity to the length of the behavioral stream and their lack of attention to the multilevel nature of the data set and underlying phenomena. We then introduce a new set of quantitative methods not constrained by these limits that provide a means of accounting for heterogeneity and structure in behavioral data. These methods include the use of the log odds ratio as an indicator of interaction pattern, random effects models as a way of specifying these multilevel relationships, and empirical Bayesian estimation methods for calculating and testing parameters within these models.

Methods Currently in Use

We first introduce some terminology. We will refer to the individual or group under study as the basic sampling unit. We will use the terms “unit” and “basic sampling unit” interchangeably throughout this paper. A basic sampling unit may be observed on several occasions, each involving a separate stream of behavior with a defined beginning and end. We refer to each such occasion as an episode. Eckerman (1993), for example, observed the same 14 pairs of toddlers interacting in free-play settings five times over a 16-month period. In our terms, each toddler pair is a basic sampling unit, and each of the five behavioral streams observed for each toddler pair is an episode, resulting in a data set with 70 separate episodes. In Eckerman’s study, each episode was limited to 16 minutes of observation, and contains a sequence of behaviors, each behavior assigned to its own category. Finally, each episode may be later broken up into two or more sub-episodes by the investigator, based on empirical or theoretical reasons. Gottman (1979), for example, observed single episodes of interaction in each of 28 couples, then broke each episode into three sub-episodes of equal length to study whether interaction patterns changed over the course of the episode.

Behavior during the episode may be shaped by things that happened earlier in the episode. Many theories hypothesize such effects, including behavioral reinforcement theories (Patterson, 1979) and theories of conflict escalation (Snyder, Edwards, McGraw, & Kilgore, 1994). To study such processes, most behavioral researchers have concentrated on the relationship between immediately antecedent behavior (A) and the immediately consequent action (C). We use the terms antecedent and consequent here simply to reflect temporal contiguity; not to imply any necessary causal relationship. Two methods have been advanced as ways of quantifying the relative strength of this relationship within a particular episode. The first involves the conditional probability, or probability across the entire episode that, when behavior A occurs, behavior C will follow. Note that this conditional probability is calculated at the level of the episode.

While some researchers have used the conditional probability as a direct measure of episode-level structure (Eckerman, 1993, for example, compares mean conditional probabilities averaged across episodes and dyads for two separate groups of toddlers), others have suggested that this is not appropriate. Bakeman and Gottman (1986, p. 149) pointed out that conditional probabilities can be strongly influenced by the simple probability of occurrence of the consequent behavior, and recommended using the binomial z score developed by Sackett (1979), with modifications recommended by Alison and Liker (1982) and Gottman (1980). The z score is a measure of the extent to which a particular observed conditional probability deviates from its expected value as based on the simple unconditional probability of the consequent behavior. Bakeman and Gottman (1986, p. 157) recommend calculating a z score for each episode, and using them as scores in standard parametric techniques such as multiple regression. (In the recent second edition, Bakeman & Gottman (1997) have now withdrawn this suggestion, based on reasons similar to those we discuss below).

More recently, Bakeman and Quera (1995) demonstrated that these statistics are quite similar to an adjusted cell residual from a two-way contingency table testing the relationship between each behavioral categories and its immediate consequent. Tables 1(a) and 1(b) present such data for one observed episode for one couple from a study of the observed interactions of 254 adult couples (Howe, 1995). Data in Table 1(a) are from the initial codes used, which unitized behavior in such a way that a particular type of behavior could be repeated more than once by an actor. In this study behaviors were also defined as states; a new behavior began when the state changed, either within an actor or as the other partner began to speak. The resulting frequencies in Table 1(b) summarize all such state transitions, either within or between partners. Since by definition a state cannot follow itself in these data, cells on the diagonal take zero values.

Following Bakeman and Quera's (1995) guidelines for analyzing data where codes cannot repeat, an adjusted cell residual can be calculated for each of the remaining 12 cells in this table, reflecting the fact that with an observational system involving four categories of behavior (in this case, two for the male partner: male negative and male positive, two for the female partner: female negative and female positive) we are able to identify 12 possible immediate antecedent-consequent relationships, since any behavior can be followed by one of three other behaviors. Note that each of these adjusted cell residuals is calculated at the level of the episode (or in this case, at the level of the couple or basic sampling unit, since each couple was observed on only one occasion).[1]

Investigators have used the adjusted cell residual (or its analogue, the z-score) as an episode-level variable interpreted as scaling the relative strength of contingency between two behaviors in a particular episode (e.g., Davis, Hops, Alport, & Sheeber, 1998). However, these statistics have important limitations that restrict their utility when they are applied to data from more than one episode.

Instability With Low Cell Counts

First, the adjusted cell residual or z-score can become quite unstable when there are few instances of a particular antecedent-consequent sequence. Figure 1 illustrates this for an adjusted cell residual based on sequences of Male Partner Negative followed by Female Partner Negative for each of the 254 couples mentioned above. The range of adjusted cell residuals is quite large for couples having the fewest instances of this sequence, and becomes much more restricted as the number of sequences increases. The effects of this property can be reduced in two ways. Investigators can drop all cases with low cell frequencies from their analyses. As an alternative, investigators could pool data across episodes or across basic sampling units into one larger contingency table to increase cell counts. Both of these options can introduce problems, however. The former may restrict the analyses to a biased subsample of cases, while the latter ignores potentially important information about between-episode or between-unit variation in antecedent-consequent patterns. In addition, as Wickens (1993) has pointed out, pooling can seriously distort findings, and can even lead to cases where associations in the pooled data are opposite to those found in each individual table.

Adjusted cell residuals or z-scores are unstable because their values are influenced by the relative frequency of the antecedent behavior. Episodes with fewer occurrences of the antecedent will be associated with greater measurement error and greater instability in the adjusted cell residual. When adjusted cell residuals or z-scores are calculated and used as predictors in parametric analyses such as regressions, it is assumed that these indicators are measured without error, an assumption that is most likely untrue. In addition, the measures from all episodes are given equal weight regardless of their accuracy. In the next section we will present a method that allows us to model and take into account this measurement error, and to give more weight to episodes with less error.

Influence of Length Of Episodes

Second, because of the way it is defined, the magnitude of the adjusted cell residual or z-score can be directly influenced by the overall length of an episode, independent of the actual relationship between an antecedent and a consequent behavior. To illustrate this, consider the simulated data where the pattern of counts is similar to that of Table 1(b), but in which each cell count is multiplied by 10, reflecting a situation where we observe an episode following the same pattern for a much longer period of time. Table 1(d) reports the adjusted cell residuals for these simulated data. In each case, the adjusted residuals are several times greater than those in Table 1(c) (in fact, they are an exact multiple of $\sqrt{10}$). This may not be a problem when data are based on episodes with equal numbers of total behaviors, but can introduce significant extraneous variability when episodes differ in length. In the couples data set we have been using as an example, each couple was observed for 15 minutes, but this resulted in episodes that ranged from 12 to 300 total behaviors.

Interpretation of Individual Cells

Third, adjusted cell residuals or z-scores can be misleading when individual cells, reflecting a particular antecedent-consequent relationship, are singled out for analysis and interpretation. These statistics use information from the entire table in computing values, and so the value for each cell is definitionally "confounded" with values from all other cells. The information on contingency in any table contains many fewer degrees of freedom than the number of cells making up the table. For example, Table 1(b) contains 16 cells but only 5 degrees of freedom. If we focus only on the subtable involving state transitions from male partner to female partner, the upper right quadrant in Table 1(a), there are four cells but only one degree of freedom. Interpretation of individual cells without taking into account the patterns in other cells can be problematic because of this.

For example, marital researchers have used the term negative reciprocity to describe the likelihood that one partner will respond negatively to a negative behavior by the other (Gottman, 1979), and the term negative reactivity to describe the likelihood that positive behavior will be suppressed following negative behavior by the other (Margolin & Wampold, 1981). It is not often noted that these two patterns are likely to be dependent, and in fact when cells from the four-cell state transition subtable are used to assess these processes, they are completely confounded, since that subtable has only one degree of freedom. This lack of independence is not just a statistical oddity, but is in fact inherent in any sequential observational data that categorizes behavior into a limited set of categories. In our example, the female partner has only two choices of response. The more she responds negatively in the face of a negative behavior from her partner, the less she responds positively. As negative reciprocity increases, so must positive responding (negative reactivity) be suppressed.

Lack of Attention to Hierarchical Structures

Finally, as we noted earlier, antecedent-consequent patterns are embedded within subepisodes, sub-episodes are embedded within episodes, and episodes may be embedded within individuals or groups. Current methods for studying antecedent-consequent patterns may be limited in their ability to unpack and make sense of this multilevel structure. Pooling data across episodes or basic sampling units eliminates information about variation between episodes or units, and this is often the variation in which we are most interested. On the other hand, using z scores or adjusted cell residuals calculated for each episode assumes these scores are measured without error or with constant error variance, an assumption that is likely erroneous.

We now introduce a different method, empirical Bayesian random effects modeling (EBREM), for studying antecedent-consequent relationships in sequential behavioral data. We apply EBREM to the log odds ratio as an indicator of contingency, which avoids some of the pitfalls of using indicators based on single cells. As we will demonstrate, EBREM not only incorporates multilevel structure explicitly in its models, but also avoids the problems of unrealistic estimates that may occur when some episodes have relatively few instances of a sequence of interest. Our work builds on and extends recent discussions of Markov models for studying variation in individual behavior chains (Gardner, 1990), and work on adjusting for between-subjects variability in contingency tables (Wickens, 1993).

Random Effects Modeling of Log Odds Ratios

We begin with some notation. Table 2 summarizes all possible two-step sequences that could reflect antecedent-consequent relationships for a set of behavior categories used to describe one complete behavioral sequence for a single episode m within the lth sampling unit. Each episode will have an $I \times J$ contingency table where the cells are filled by frequency counts for sequences beginning with one of the I codes (antecedent) and ending with one of the J behaviors (consequent). Let n_{ijml} be a frequency count for the sequence beginning with behavior i and ending with behavior j for the mth episode and lth sampling unit, where $i = 1, \cdots, I, \quad j = 1, \cdots, J$, $m = 1, \cdots, K$, and $l = 1, \cdots, L$ where K is the total number of episodes per sampling unit and L is the total number of sampling units in the study. In many cases $I = J$, since the same set of behavioral categories will be considered as both antecedents and consequents.

Use of the Log Odds Ratio

Upton (1982) identified 22 measures of association that have been developed for summarizing data in contingency tables. The adjusted cell residual that we have been discussing has two virtues as a measure of association between antecedent and consequent: it is not affected by rates of the antecedent behavior, and it defines strength of association in terms of how much the cell frequency reflecting a particular antecedent-consequent association deviates from the expected or "chance" value. We would like to retain these characteristics, but also have an indicator that is not sensitive to the total number of behaviors in the sequence, and is tractable when involved in more complex models.

We begin with the odds for a particular cell, defined as the conditional probability for the consequent given the antecedent, divided by the conditional probability for all other consequents given that antecedent. Using our notation, the formula for the odds of cell ij for the mth episode and lth sampling unit is given by: $P_{ml}(j|i)/(1 - P_{ml}(j|i))$, where $P_{ml}(j|i)$ is the probability of occurrence of event j as consequent behavior given that event i happened as antecedent behavior.

The odds meets two of our criteria: it is not affected by either the antecedent marginal frequency or the total table frequency. However, the simple odds can be affected by the marginal frequency of the consequent. To remedy this, we move to an odds ratio, which compares the odds that a particular consequent will follow the antecedent to the odds that the consequent will follow all other relevant antecedents.

What do we mean by relevant antecedents? We use this term because antecedent-consequent relationships in any study can be of different logical forms, and the appropriate set of antecedents to be used here will depend on the substantive questions to be studied.

The example data in Table 1(a) include three different forms of antecedent-consequent relationships involving (1) stability within actor, (2) state change within actor, and (3) state change between actors. Suppose we are interested in the pattern of state changes from male partner to female partner, and we wish to determine the pattern and strength of this association using the odds ratio, with particular reference to the cell reflecting Male Negative followed by Female Negative. We could base this ratio on the entire 4×4 table, defining Female Positive, Male Negative, and Male Positive all as relevant antecedents. The resultant odds ratio would tell us how strongly this sequence occurred in comparison to those following all other antecedents, including the Male Negative followed by Male Negative sequence. The data in Table 1(a) however strongly suggest that the pattern of self-stability reflected in the Male Negative to Male Negative sequence is very strong, and it probably reflects a very different process than the cross-actor association in which we are interested. If this is the case, then including Male Negative as a relevant antecedent would greatly reduce the odds ratio, inappropriately comparing the cross-actor sequence of interest to a within-actor sequence that is influenced by a fundamentally different process.

A second option would be to transform our dataset to one that includes only state changes. In this case, no behavior may follow itself, and cell frequencies reflect only those points of transition from one state to another. Data in Table 1(b) are based on the same observations as those in Table 1(a), but include only state transitions. While we might use the entire table to calculate our odds ratio, the pattern of frequencies in the within-actor blocks appears very different from those in the between-actor blocks, with very few instances of within-actor transitions occurring once we have eliminated those involving state stability. Again, including the within-actor transitions would seem to be inappropriate.

A third option, and one that we would advocate for this particular example, would define as relevant antecedents only those antecedents involved in cross-actor transitions. For these data, the odds ratio of interest would involve only one other antecedent, Male

Positive, and would be based on the four cells in the upper right quadrant of Table 1(b). The observed odds ratio would be calculated as $(30/27)/(16/42) = 2.92$ and the log odds ratio as $log(2.92) = 1.07$.[2] Note that the log odds ratio uses information from all four cells in the subtable in its calculation.

Table 3 provides the observed log odds ratios for the subtable of interest. The four log odds ratios are clearly not independent of one another, and in fact are either equivalent or exact inverses of one another. This reflects that fact that any 2×2 table has available only one degree of freedom for testing level of association. For this particular couple, the log odds ratio reflects a combined pattern of higher negative reciprocity and positive reciprocity as well as higher negative and positive reactivity (suppression of positive following negative, or negative following positive).

These odds ratios are based on data from a single couple, and in this data set may be estimated for each of the $L = 254$ episodes, or couples (since there is only one episode per couple). The distribution of a particular odds ratio across the population of episodes is positively skewed and nonnormal. Figure 2(a) depicts the histogram across couples for the odds ratio of the transition from Male Partner Negative to Female Partner Negative, illustrating this positive skewness. Taking the natural logarithm of the odds ratio makes the distribution symmetric, as is evident in Figure 2(b). The distributional advantages of improved symmetry and approximate normality lead us to prefer using the log odds ratio as an indicator of a strength of a particular antecedent-consequent relationship over other candidates, such as the simple conditional probability.

The log odds ratio has other advantages over the use of the conditional probability. Log odds ratios can range anywhere from plus to minus infinity so, unlike probabilities which range between zero and one, log odds ratio models do not lead to any range restrictions. This means that mathematical models using the log odds ratio are simpler and more tractable (although there have been some attempts to develop models for conditional probability indexes in other cases involving data with multilevel structure: e.g., Dersimonian & Laird, 1986; Wickens, 1993). Modeling on the log odds ratio scale has the advantage that most effects can enter as additive terms. For example, a test of whether two conditional probabilities are equal can be reexpressed in terms of the two associated log odds being equal to one another. An equivalent way of expressing this is that the log odds ratio, or the difference in the two log odds, is equal to zero. The value of zero is a natural center point for modeling log odds ratios. A log odds ratio of zero is equivalent to the independence or absence of association between antecedent and consequent.

Modeling and Estimation of the Log Odds Ratio

In the remainder of this paper, we advance a general modeling framework for using log odds ratios as indicators of contingency in behavioral data. This modeling framework is composed of two distinct but integrated components: the estimation of true log odds ratios from the contingency table, and the estimation of these true log odds ratios based on multilevel models.

Here we discuss two ways that log odds ratios can be estimated: direct calculation, and empirical Bayesian modeling. First, log odds ratios for each episode can be calculated directly from the data for that episode. We term this the observed log odds ratio. To reduce the complexity of our notation, let the total number of episodes in our data set be M, where $M = KL$ (episodes per sampling unit times number of sampling units). When only one episode is observed per sampling unit, $M = L$. Let a_m be a frequency count for a particular behavior followed by another behavior of interest (e.g., Male Partner Negative followed by Female Partner Negative) from the mth episode and let b_m be the frequency count for the male partner negative followed by female partner positive, c_m be a frequency count for the male partner positive followed by female partner negative and d_m be a frequency count for male partner followed by female positive. Thus, the observed log odds ratio can be calculated as $\log(a_m d_m/(b_m c_m))$. This expression works well when a, b, c and d are large, but in small samples there is measurable bias in the estimate of the log odds ratio. To reduce bias one often uses a slightly modified estimate of the log odds ratio as

$$Y_m = \log((a_m + 1/2)(d_m + 1/2)/[(b_m + 1/2)(c_m + 1/2)]) \tag{1}$$

which works especially well even if cell counts a, b, c or d are zero. So-called exact methods are also available (e.g., MIXOR: Hedeker & Gibbons, 1994), but in our examples the results obtained from special programs such as MIXOR are not likely to have much effect on our estimates. We have chosen to model the log odds ratio directly, but the same model could be expressed in terms of the logit. Log odds ratio and logit (logistic) modeling are directly related as the log odds ratio is the difference between the logits of conditional probabilities of one partner responding negatively given that the other partner initiated with either negative or positive behavior. We may also calculate the variance of Y_m, S_m^2, using the formula

$$S_m^2 = 1/(a_m + 1/2) + 1/(b_m + 1/2) + 1/(c_m + 1/2) + 1/(d_m + 1/2). \tag{2}$$

Note that in our example each observed log odds ratio appearing in Table 3 has a separate variance. This variance reflects the fact that the precision of the observed log odds ratio as an estimate of the true log odds ratio increases as the number of behaviors in an episode increases. Observed log odds ratios based on shorter sequences of interaction will typically have larger variances than observed log odds ratios based on longer sequences. This characteristic of observed log odds ratios highlights an important limitation. If we have a sample of episodes that vary in overall length, then there will be substantial variation in how well the observed log odds ratio reflects the true log odds ratio in each episode. However, if we include these observed log odds ratios as independent or dependent variables in standard statistical analyses, measurement error is assumed to be equal across all observed episodes, an assumption which does not in fact hold. This situation may occur when episodes vary in length, but it can also occur when our sample of episodes shows no variation in overall length, but great variation in various types of transitions. For example, we could find substantial variation in the number of Husband-to-Wife state transitions that occur in each episode, even though overall episode length is fairly similar.

Empirical Bayesian Formulation for the Log Odds Ratio

Empirical Bayesian techniques, a second method of estimation, have the advantage of explicitly including information about variations in precision of estimation in the modeling of true log odds ratios. They do so in part by placing the log odds ratio in a multilevel framework. The multilevel model is becoming increasingly well-known in the social sciences (Bryk & Raudenbush, 1992; Hedeker & Gibbons, 1994), and is useful for the analysis of data with hierarchical structure, as is the case here.

First, we specify the random variation in log odds ratios across episodes. Let us begin by focusing on one log odds ratio from a set of log odds ratios that might be used to characterize an entire contingency table. Let Y_m be the observed log odds ratio for the *mth* episode. Under the simplest random effects model, it is assumed that the observed log odds ratio is modeled as the sum of the true log odds ratio and within episode error,

$$Y_m = \theta_m + E_m \tag{3}$$

where the within episode error variable, E_m, is generated from a normal distribution with mean zero and variance σ_m^2. The normal probability density for this distribution is given by $f(Y_m|\theta_m, \sigma_m^2) = (2\pi\sigma_m^2)^{-1/2} exp\{-1/2(Y_m - \theta_m)^2/\sigma_m^2\}$, and θ_m is the true log odds ratio for the *mth* episode. We use the term "true" log odds ratio to refer to the unobserved log odds ratio, θ_m. This usage is similar to that in factor analysis and other measurement models using latent variables. As we shall see later, it does not mean that E_m accounts for all measurement error in the observed log odds ratios.

In this specification, the true log odds ratios are themselves random variables generated from a superpopulation that has mean μ and variance τ^2. By incorporating μ and τ^2 in the random effects model, we can then study variables that may be sources of variation in log odds ratios across episodes.

We specify the second level of this model by allowing the true log odds ratio for each episode to be modeled as the sum of the common log odds ratio, μ, and the between episode variability,

$$\theta_m = \mu + V_m \tag{4}$$

where the variation among episodes, V_m, has a normal distribution with mean zero and variance τ^2. That is, the conditional density for the parameter θ_m is given by $f(\theta_m|\mu, \tau^2) = (2\pi\tau^2)^{-1/2} exp\{-1/2(\theta_m - \mu)^2/\tau^2\}$. When τ^2 is zero, the model collapses to a simple fixed effect model, and the log odds ratio is constant for all episodes.

By combining equations (3) and (4) via Bayes Theorem (Carlin & Louis, 2000), the conditional density of the true log odds ratio is given by

$$\begin{aligned} f(\theta_m|Y_m, \mu, \tau^2, \sigma_m^2) &= const \;\; f(Y_m|\theta_m, \sigma_m^2) f(\theta_m|\mu, \tau^2) \\ &= const \;\; (\tau\sigma_m)^{-1} exp\{-1/2[(Y_m - \theta_m)^2/\sigma_m^2 + (\theta_m - \mu)^2/\tau^2]\}, \quad (5) \end{aligned}$$

where *const* is a proportionality constant. Algebraic recombination on (5) shows that the true log odds ratios, θ_m, have normal distributions with means, $\theta_m^* = (1 - \alpha_m)Y_m + \alpha_m\mu$, where $\alpha_m = \sigma_m^2/(\sigma_m^2 + \tau^2)$. These means are weighted combinations of the observed log odds ratios Y_m and the overall mean of log odds ratios μ. The variances of the true log odds ratios (θ_m) are also obtained as $\sigma_m^2(1 - \alpha_m)$.

From the above results it can be seen that θ_m^* is a weighted average of μ and Y_m (the observed log odds ratio). The weights, α_m, depend on the relative sizes of the variance between episodes (τ^2) and the within-episode variance σ_m^2. As τ^2/σ_m^2 becomes smaller (meaning less variability between episodes), more weight is given to μ. This acts to improve the precision in estimation of each true log odds ratio by borrowing information from all episodes (since observations from all episodes are used to estimate μ and τ^2). In particular this helps stabilize estimates for cases with relatively small cell counts, which have greater within-episode variance.

Empirical Bayesian Estimation and Inference

Provided that μ, σ_m^2, and τ^2 are known, θ_m^* is the Bayes estimate of θ_m (the true log-odds) since it is the mean of the conditional posterior distribution for the θ_m. But, in most situations, these parameters are unknown. They have to be estimated from the data. When they are replaced by their estimates, θ_m^* is referred to as the empirical Bayes estimate of θ_m. The calculation of maximum likelihood estimates of μ and τ^2 can be performed using iterative procedures on the following formulae.

$$\hat{\mu} = \sum w_m Y_m / \sum w_m, \tag{6}$$

$$\hat{\tau}^2 = \sum (w_m^2 [(Y_m - \hat{\mu})^2 - S_m^2]) / \sum w_m^2, \tag{7}$$

where $w_m = 1/(S_m^2 + \hat{\tau}^2)$ which is the weight that incorporates random effects variance estimate $\hat{\tau}^2$; Y_m is the observed log odds ratio, and the summation is over the L episodes. When the quantity on the right side of equation (7) is negative $\hat{\tau}^2$ is set to zero since variances must be nonnegative. The within episode variance σ_m^2 is estimated by S_m^2 (defined earlier). Because all the observed data Y are used to find estimates of μ and τ^2, the estimates are improved by *borrowing strength* from the other observations to make inferences about a particular θ.

A generic Splus algorithm to compute the above estimates is available in Appendix A. We note that some of the models (see equations 8 and 11) we examined in this paper are similar but not identical to ones available in standard software packages such as HLM and SAS Proc Mixed. Even though HLM readily handles known variances at level-1 (Bryk & Raudenbush, 1992, p. 172), it does not readily allow for special structures involving the means and variances. SAS Proc Mixed has somewhat more flexible variance structures but still cannot currently fit all of our models. We note that both of these programs could be modified to fit all the models we describe, if they incorporated Lagrange multiplier techniques that are routine in structural equation modeling (Sörbom, 1989).

Hierarchical Extensions of Random Effects Models

In this section we turn more fully to the issue of hierarchical structure in behavioral data. Many of the questions posed by observational researchers reflect this inherent hierarchical structure. If one action prompts a specific response during part of an episode, does its effect vary across other parts of the episode (is there within-episode variation)? If one observes episodes of interaction both before and after family therapy, do contingencies between behaviors change across these episodes (is there cross-episode variation within unit)?

Here we extend our application of empirical Bayesian estimation to more complex hierarchical models that can shed light on such questions. The basic random effects model we have developed here can serve as the foundation for a series of more complicated models used to study various types of structure in antecedent-consequent relationships. First, investigators may wish to determine whether the strength of the antecedent-consequent association remains constant throughout individual episodes, a condition called *stationarity.* Next, investigators are usually interested in whether this association varies systematically across a sample of episodes, or whether the same antecedent-consequent association obtains for all episodes, a condition termed *homogeneity.* If episodes show heterogeneity, the source of this heterogeneity may involve within-unit variation, as when multiple episodes are collected for each sampling unit, and antecedent-consequent relationships vary across those episodes within each sampling unit. Finally, investigators are usually interested not only in determining whether there is variation in antecedent-consequent associations, but also in what factors might account for or result from such variation. In the following sections we present models for studying each of these types of structure: stationarity, homogeneity, within-unit variation, and covariation with other factors.

Testing for Significant Variation Within Episodes

The strength of contingency between two behaviors may be constant throughout an episode. For example, we might hypothesize that couples vary in how reactive each partner is to negative statements made by the other, and that this reactivity has a consistent and stable effect for each couple throughout the episode of interaction. However, it may also be the case that reactivity itself changes across the course of the episode, and is stronger in earlier sub-episodes than in later ones. Couples might habituate to each other′s negative comments, or become even more sensitive to them as the discussion continues. Most investigators assume the contingencies they study are stationary within episode, although a few have hypothesized and studied systematic variation within episodes (Gottman, 1979).

The random-effects model allows us to test for the presence of such variation, or nonstationarity. There are three approaches to specifying a model to provide such a test. First, the investigator can arbitrarily break each sequence into halves or thirds, and test for significant variation across these sub-episodes. This method has been recommended in the case where the investigator considers nonstationarity as a nuisance factor that must be ruled out in order for the assumptions of other analyses to hold (Gottman & Roy, 1990).

As an alternative, an investigator may have theoretical reasons to believe that contingencies will vary for certain portions of an episode as compared to others. This would be the case when experimental designs are used that introduce contextual changes in the midst of an episode. Minuchin, Rosman, and Baker (1978), for example, observed couples discussing an area of disagreement when their child was not present, and then asked the parents to continue discussing the topic when the child was brought into the room. With this experimental design each couple′s interaction could be broken into two subepisodes, one prior to and one following the entrance of the child.

Finally, it is possible to model continuous variation in contingencies across the course of an episode, without requiring the investigator to identify points of qualitative change.

Such models could reflect for example, an escalation, de-escalation, or nonlinear pattern of a contingent response over the course of the episode. We know of no observational studies that have as yet attempted to test for such patterns of change, however.

Here we provide an example to illustrate the first approach. In this case, each episode will be broken into two subepisodes having equal numbers of behaviors. We begin by respecifying the random effects model in the following way.

Level 1:

$$\begin{aligned} Y_{m_1} &= \theta_{m_1} + E_{m_1}, \quad \textit{and} \\ Y_{m_2} &= \theta_{m_2} + E_{m_2}, \end{aligned} \tag{8}$$

where Y_{m_1} and Y_{m_2} are the observed log odds ratios for the first and second subepisodes, θ_{m_1} and θ_{m_2} are the true log odds ratios for these subepisodes, and E_{m_1} and E_{m_2} are distributed normally with mean zero and variances $\sigma^2_{m_1}$ and $\sigma^2_{m_2}$.

If these data are stationary, then the true log odds ratio in the first subepisode should equal the true log odds ratio in the second subepisode. More formally, if we assume stationarity within episodes, but allow the log odds ratios to vary across episodes, we have

Level 2:

$$\theta_{m_1} = \theta_{m_2}. \tag{9}$$

We can systematically examine deviations from stationarity by considering level 2 models which incorporate these constraints. A general level 2 model[3] that allows us to test for such deviations is specified as

Level 2:

$$\begin{aligned} \theta_{m_1} &= \mu_1 + \delta_m + V_{m_1}, \\ \theta_{m_2} &= \mu_2 + \delta_m + V_{m_2}. \end{aligned} \tag{10}$$

In these equations δ_m is a random variable that carries information about the portion of the log odds ratio that is the same for both subepisodes, and has mean zero and variance ψ^2. The parameters μ_1 and μ_2 carry information concerning how the mean values of the log odds ratios may vary across the two subepisodes. V_{m_1} and V_{m_2} are independent normal errors with mean zero and variances τ_1^2 and τ_2^2, respectively. τ_1^2 and τ_2^2 carry information about "subepisode-level" variation in log odds ratios. Based on equations (8) and (10), the joint distribution of Y_{m_1} and Y_{m_2} becomes bivariate normal with means μ_1 and μ_2, and variance-covariance matrix given as

$$\begin{pmatrix} \sigma_{m_1}^2 + \psi^2 + \tau_1^2 & \psi^2 \\ \psi^2 & \sigma_{m_2}^2 + \psi^2 + \tau_2^2 \end{pmatrix}.$$

It follows from the above matrix that marginally Y_{m_1} and Y_{m_2} have variances $\tau_1^2 + \psi^2 + S_{m_1}^2$ and $\tau_2^2 + \psi^2 + S_{m_2}^2$, and covariance ψ^2. The error variance, σ_m^2, changes from episode to episode.

Note that the overall test of stationarity requires $\mu_1 = \mu_2$ and $\tau_1^2 = \tau_2^2 = 0$. The true log odds ratios for the two subepisodes will be equal only if all these conditions hold. These equations also specify different aspects of deviation from stationarity, and these can be tested separately. To test where there is a "drift" in the mean log odds ratio from time period 1 to time period 2 for the entire sample, we can test $H_0 : \mu_1 = \mu_2$. To test whether the extra variation in log odds ratios within episodes is the same at the two points in time we can test $H_0 : \tau_1^2 = \tau_2^2$. Finally a test of $H_0 : \tau_1^2 = \tau_2^2 = 0$ examines whether there is any extra variation in log odds ratios beyond a factor common to the entire episode.

To illustrate, we tested for departures in stationarity for the contingency data involving transitions from Male Partner to Female Partner, based on the two-by-two subtable used as an example earlier. For this analysis, each episode for each couple was divided into two subepisodes having equal numbers of behaviors, randomly assigning the middle behavior to one subepisode in cases where the sequence involved an odd number of behaviors. The analysis was conducted using the log odds ratio reflecting Male Partner Negativity followed by Female Partner Negativity. Identical conclusions would have been obtained using any of the other three cells.

We used iterative maximum likelihood methods to compute the parameter estimates of the proposed models using a developmental version of the Mplus program (Muthén & Muthén, 1998). Copies of the input used for the analyses in this paper can be obtained from the first author and will be available on the Mplus website (*www.StatModel.com*) when the new version (2.03) is released. As stated earlier, currently the HLM and SAS statistical software packages lack flexibility in incorporating special covariance structures such as ours.

In all models heterogeneity was allowed (i.e., ψ is unrestricted). That is, log odds ratios were allowed to vary across episodes. We discuss tests of heterogeneity in the next section. The results are given in Table 4.

Model 1 in Table 4(a) is labeled as unrestricted, meaning that no stationarity is assumed. Model 2 represents the most restricted form of stationarity, including all three requirements for stationarity to hold. Since these models are hierarchically nested, minus twice the difference in their log likelihoods, symbolized as G^2, is approximately distributed as chi-squared with degrees of freedom equal to the difference between the numbers of parameters specified under the restricted model and the unrestricted model. We can therefore test whether the model restrictions imposed in Model 2 lead to a significantly poorer fit than that of the unrestricted model. In this case the change in log likelihood is significant ($G^2(3) = 12.97$, $p < 0.005$), and we must reject the hypothesis that these data are stationary.

Models 3 and 4 allow for separate tests of the two different components of stationarity. Model 3 restricts the mean log odds ratio for the two subepisodes to be equal, but allows their variances to differ, while model 4 allows the mean log odds ratio for the two subepisodes to differ, but restricts their variances to zero. Model 3 does not fit as well as the unrestricted model, suggesting that the mean log odds ratios do differ for the two subepisodes. The constraints in Model 4 lead to a tendency toward poorer fit, but only weakly.

Table 5 reports point estimates and 95% confidence intervals for parameters in the unrestricted model. Both estimates of mean log odds ratios differ significantly from zero, indicating that the association between Male Partner and Female Partner behavior is significantly strong. From model 3, we have evidence that there is a significant increase in the mean log odds ratio from the first to the second subepisodes. This suggests that, on overage, female partners become more likely to "mirror" the behavior of their male partner (negative following negative, positive following positive), and less likely to change the valence of the interaction (reduced rates of negative following positive, or positive following negative), over the course of the interaction.

How are we to interpret the variance parameters, τ_1^2 and τ_2^2? Such variation could have two sources. It might reflect systematic changes in the log odds ratio across the two subepisodes, such that the strength of those changes differed for different episodes. This variation could also include measurement error. As we noted earlier, while the first level component of this model accounts for sources of measurement error related to the length of a sequence, it does not model measurement error from other sources, such as less-than-perfect reliability in the behavioral coding system. While we cannot partition variation due to systematic nonstationarity from that due to measurement error in this model, we can use the model as a basis for identifying potential sources of systematic nonstationarity, as we describe later.

In these data, there appears to be surprisingly little variation at this level. The overall test of Model 4 does not reach significance, and the confidence interval for τ_1^2 overlaps with zero. This suggests that much of the error of measurement was accounted for by variations in sequence length that influence E_m, which is taken into account in the Level 1 model. It also suggests that systematic nonstationarity at this level is likely to be weak or nonexistent.

Testing for Significant Variation Across Episodes

Since the random effects model allows us to partition the variation in the estimated log odds ratios into that due to within-episode factors and that due to between-episode factors, we can now use the log odds ratio estimator to test whether there is significant variation in the strength of a particular pattern of transitions across the entire sample; that is, whether all episodes have essentially the same sequential structure. In other words, we can test whether there is significant heterogeneity among episodes. Specifically, $H_0 : \theta_1 = \theta_2 = \cdots = \theta_L$.

When data meet the criteria for stationarity, it is reasonable to recombine subepisodes into single long episodes, and use a simpler method for testing whether the episode-level log odds ratios vary across the sample of episodes. Here we can use the most popular test for heterogeneity used in meta-analysis (Hedges & Olkin, 1985; Takkouche, Cadarso-Suárez, & Spiegelman, 1999). To employ this test, each log odds ratio is assigned a weight (W_m) based on the inverse of its variance ($W_m = 1/S_m^2$). For each episode, the weight is multiplied by the natural log of odds (Y_m) to give a summary measure (W_mY_m). A pooled summary is calculated by dividing the sum of the summary measures over all L episodes by the sum of the weights ($\bar{Y} = \sum W_mY_m / \sum W_m$). Now, heterogeneity among episodes can be assessed by using the chi-square statistic ($\chi^2 = \sum(W_m(Y_m - \bar{Y})^2)$ with $L-1$ degrees of freedom, where L is the total number of episodes. Large values of the test statistic provide evidence for rejecting the homogeneity hypothesis. For our example of log odds ratios for Male Partner Negativity followed by Female Partner Negativity, the value of the chi-square statistic is 475.20 with 253 degrees of freedom and p-value less than .001. Thus, there is strong evidence that there is heterogeneity among episodes.

If data do not meet the criteria for stationarity however, as is the case in our example, this simpler test of homogeneity may provide misleading results. In this case, we can return to a full two-level modeling approach that incorporates parameters for testing heterogeneity at both the subepisode and episode levels. In the two-level model specified in equations (8) and (10), δ_m is an indicator of the episode-level log odds ratio, and its variance ψ^2 measures how much variation there is in episode-level log odds ratios across the sample of episodes. In addition, τ_1^2 and τ_2^2 reflect the presence of variation in the log odds ratios for the first and the second subepisodes above and beyond that found at the episode level. An overall test of homogeneity in the presence of nonstationarity would involve setting these three parameters equal to zero, and testing whether this model fits the data as well as the unrestricted model (Model 1 versus Model 5 in Table 4(b)).

Table 4(b) reports results for our example. Model 5, which incorporates these restrictions, significantly reduces fit in comparison to the unrestricted model, providing strong evidence for heterogeneity in the presence of nonstationarity. Model 6 suggests that there is significant episode-level heterogeneity to be explained ($G^2(1) = 5.77, p < .016$) in the presence of nonstationarity.

In addition to this episode-level heterogeneity, the point estimate and confidence intervals for τ_2^2 reported in Table 5 suggest that there is significant subepisode-level heterogeneity for both subepisodes, and this may include systematic variation that could also be explained, although Model 3 results indicate this variation is small.

Testing for Variation Between Episodes Within Sampling Units

If we find evidence for significant heterogeneity in the estimated log odds ratios across the sample of episodes, we can then develop and test models that account for this heterogeneity. If we have collected data on only one episode per sampling unit, we can proceed directly to such models. If we have collected data for more than one episode per basic sampling unit however, then it is possible that some of this heterogeneity is due to variation within each basic sampling unit. The meaning of such variation will of course depend on how the episodes were sampled. For example, we might wish to study whether certain antecedent-consequent relationships are altered after some form of intervention. For example, Revenstorf, Hahlweg, Schindler, and Vogel (1984) observed and coded couples behavior before and after participation in marital therapy. They reported conditional probabilities suggesting that problem escalation, reflecting high rates of responding negatively to problem descriptions by the other, were lower after therapy than before. While they reported results based on simple inspection of conditional probabilities from these two sets of episodes for each sampling unit, a random-effects model could be used to test whether there was in fact significant variation across episodes that might reflect the salutary effects of their intervention.

The model for specifying cross-episode variation when episodes are nested within basic sampling unit is, in fact, conceptually identical to the model we specified earlier for studying stationarity by breaking episodes into sub-episodes, and testing variation across sub-episodes as they are nested within overall episode. The model can be specified as follows for a study with three episodes ($k = 1...3$) for each sampling unit:

Level 1:

$$\begin{aligned} Y_{1l} &= \theta_{1l} + E_{1l} \\ Y_{2l} &= \theta_{2l} + E_{2l} \\ Y_{3l} &= \theta_{3l} + E_{3l} \end{aligned} \tag{11}$$

Level 2:

$$\begin{aligned} \theta_{1l} &= \mu_1 + \delta_l + V_{1l} \\ \theta_{2l} &= \mu_2 + \delta_l + V_{2l} \\ \theta_{3l} &= \mu_3 + \delta_l + V_{3l} \end{aligned} \tag{12}$$

where, Y_{kl} is observed log odds ratio for the kth episode within the lth sampling unit ($k = 1, 2, 3$) and E_{kl} is the error term distributed as normal with mean zero and variance σ^2_{kl}, θ_{kl} is the true log odds ratio and V_{kl} is the random effect associated with the kth episode within the lth unit and is distributed as normal with mean zero and variance τ^2_k. Here μ_1, μ_2, and

μ_3 carry information about the mean log odds ratio for each episode across the set of basic sampling units, and δ_l carries information about the log odds ratio for each unit as a whole. Thus, the joint marginal distribution of Y_{1l}, Y_{2l}, and Y_{3l} is multivariate normal with means μ_1, μ_2, and μ_3, respectively, and variance-covariance matrix

$$\begin{pmatrix} \sigma_{1l}^2 + \psi^2 + \tau_1^2 & \psi^2 & \psi^2 \\ \psi^2 & \sigma_{2l}^2 + \psi^2 + \tau_2^2 & \psi^2 \\ \psi^2 & \psi^2 & \sigma_{3l}^2 + \psi^2 + \tau_3^2 \end{pmatrix}.$$

Note that this model allows for correlations among episodes within each unit, by including in the level 2 model the common random-effects parameter δ_l whose variance is ψ^2.

The data we are using for illustration do not include multiple episodes of observation for each sampling unit, so we are unable to present an example of these tests. However, the estimation procedures and logic of interpretation are strictly analogous to those used in the tests of stationarity and homogeneity presented earlier. For example, equation sets 11 and 12 are simply trivariate extensions of the bivariate model described earlier in equation sets 8 and 10. This hierarchical approach would also allow for specification and testing of more complex models that test stationarity within each of several episodes nested in turn within basic sampling units.

Introducing Covariates Into the Model

Assuming we find evidence for significant heterogeneity in log odds ratios, the random effects model allows us to introduce covariates measured at various levels into the model, to determine how they might be related to the strength of a particular antecedent-consequent relationship.[4] In the Revenstorf et al. study described earlier, participation in the intervention could serve as a covariate measured at the level of episode, distinguishing episodes prior to intervention from those following. An analysis of covariance approach, where pre-intervention contingent response is treated as a predictor of post-intervention response, allows one to assess changes in contingent response over time. Covariates could also be introduced at the level of the basic sampling unit, which was the couple in the Revenstorf et al. study. There are other situations where interactions, say between sampling unit characteristics and intervention status, could be used to examine variation in intervention impact (Brown & Liao, 1999).

First we consider the simpler case where earlier tests provide evidence for heterogeneity in the presence of stationarity. To incorporate episode-specific covariates in equation (4) so that heterogeneity among units can be accounted for, we model the log odds ratio as

Level 1:

$$Y_m = \theta_m + E_m, \tag{13}$$

Level 2:

$$\theta_m = \mu + \beta X_m + V_m, \tag{14}$$

where X_m is a covariate for the mth episode and β is a fixed effect coefficient. E_m and V_m have the same definitions and distributional properties of those given in equations (3) and (4).

To illustrate, we now introduce an episode-level covariate to test its relationship with the estimated log odds ratio index of the strength of association between antecedent Male Partner and consequent Female Partner behavior. We chose an index of perceived relationship adjustment, based on an average score combining both partners$'$ self-report ratings on the Dyadic Adjustment Scale (Spanier, 1976), a commonly used measure that includes questions concerning perceptions of conflict, satisfaction, support, and relationship stability. The parameter (β) estimating the relationship between this episode-level variable and the behavioral contingency of interest was estimated as .174 with standard error 0.062. There is thus a significant relationship between the log odds ratio and the index of perceived relationship adjustment. This suggests that the higher the perceived relationship adjustment, the more the Female Partner "mirrors" the Male Partner, and the less she switches to the opposite valence. (Again, we interpret all four cells in the relevant subtable together, since their log odds ratios are completely redundant.) The scatter plot of the empirical Bayes estimates of the true log odds ratios and the index of perceived relationship adjustment is depicted in Figure 3.

When preliminary models indicate heterogeneity in the presence of nonstationarity, as is the case in our example, a more complex model is required. In this case, to incorporate episode-specific covariates in the random effects model we model the log odds ratio as

Level 1:

$$\begin{aligned} Y_{m_1} &= \theta_{m_1} + E_{m_1} \\ Y_{m_2} &= \theta_{m_2} + E_{m_2} \end{aligned} \tag{15}$$

Level 2:

$$\begin{aligned} \theta_{m_1} &= \mu_1 + \beta_{11} X_m + \delta_m + V_{m_1} \\ \theta_{m_2} &= \mu_2 + \beta_{12} X_m + \delta_m + V_{m_2}, \end{aligned} \tag{16}$$

where X_m is a covariate for the mth episode and β_{11} and β_{12} are fixed-effect coefficients representing separate effects for the covariate on the log odds ratios for the two different subepisodes. E_{m_1}, E_{m_2}, δ_m, V_{m_1}, and V_{m_2} have the same definitions and distributional properties of those given in equations (8) and (10).

Table 6 presents results for two models: the unrestricted model, which allows the effects of the covariate to differ for the two subepisodes, and a model that restricts the effects of the covariate to be equal across the two subepisodes. Chi-square indexes of model fit are almost equal, suggesting that the covariate is having identical effects in the first and second subepisodes. Given this result, it is not surprising that parameter estimates for the covariates are almost identical to those produced by the models presented earlier that assume stationarity.

Discussion

In this paper we have reviewed common methods for measuring strength of contingency between two behaviors in a behavioral sequence, and have pointed out a number of limitations with these approaches. We then presented a new approach employing empirical Bayes estimation in the context of hierarchical or multilevel modeling, an approach not constrained by these limitations. Here we take up three general issues concerning this new approach: how to organize a set of analyses by following a general strategy; what some of the limits are of this new approach; and what further developments in quantitative method would prove useful.

Our examples in this paper followed a stepwise strategy based on two traditions: Markov modeling (Gottman & Roy, 1990) and hierarchical modeling (Bryk & Raudenbush, 1992). From the Markov modeling tradition we took the concept of stationarity, the idea that we need to understand whether the process that leads to contingencies among behaviors in a sequence is constant across the course of an episode, or varies over the episode. We recommend testing stationarity as a first step in any hierarchical analysis of contingency. The investigator can divide episodes into subepisodes based on theoretical, empirical, or experimental design considerations, or can simply split the episode in two or more to provide a test of strong nonstationarity, as in our example.

The presence or absence of nonstationarity will dictate how to proceed. If there is no evidence of nonstationarity, the investigator can proceed with simpler tests of homogeneity based on entire episodes (after pooling data across sub-episodes); otherwise, multilevel models allow for studying homogeneity in the presence of nonstationarity.

We recommend that tests of homogeneity proceed in stages, moving up to higher levels one at a time. First, test for homogeneity across all episodes. If homogeneity is found, some have suggested the analysis be considered complete, since the data fit a fixed effect model, where there is no evidence that strength of contingency varies across episodes. In this case the fixed effects model will provide a point estimate and confidence interval for the strength of contingency across all episodes. However, it is possible that, while the null hypothesis of no heterogeneity cannot be rejected based on these tests, further analyses can find significant relationships between covariates and the latent log odds ratios. M. Stoolmiller (personal communication, October 2, 2000) has suggested that such findings may be taken as evidence against the null hypothesis of homogeneity. In such cases, the amount of variation in the latent log odds ratios may be small, but still of theoretical interest.

If the data are found to be heterogeneous, the next step involves identifying at what levels that heterogeneity occurs. If the design included observation of multiple episodes for each basic sampling unit (such as a couple or a family), we can test whether there is significant heterogeneity both within and between sampling units. If there is no evidence of heterogeneity across episodes within sampling units, the investigator is justified in pooling data from within-unit episodes into a single contingency table for each sampling unit. If however there is evidence of heterogeneity between episodes within sampling unit, more complex models will be needed that incorporate this source of variation, even if the investigator may have no substantive interest in studying within-unit variation.[5]

Most investigators will be interested in identifying and studying between-unit variation in contingency. If prior analyses show no evidence of nonstationarity or within-unit heterogeneity, subepisodes can be recombined into episodes, within-unit episodes can be pooled together, and covariates can be used in a simple two level model to study sources of between-unit heterogeneity. A major strength of EBREM lies in its capacity to provide evidence that these simplifying assumptions are appropriate, and to provide methods for studying variation in contingency even when they are not. Even when the investigator has no substantive interest in nonstationarity or within-unit heterogeneity, incorporating both sources of variation into the model will not only provide for a more accurate assessment of unit-level variation, but will allow for more precise estimates of parameters.

The EBREM approach presented here also has several limitations. The models developed here focus on only one log odds ratio at a time. Observational systems with more than two categories of antecedent or consequent behavior will lead to contingency tables with more than one degree of freedom, and will reflect several possible antecedent-consequent patterns, which may or may not be independent of one another. EBREM approaches need to be extended to such situations, with particular attention to patterns of relationship among different antecedent-consequent associations. This will most likely require multivariate approaches that model more than one contingency simultaneously, to allow for correlation among log odds ratios.

In this paper we have only focused on consequents that immediately follow antecedents, reflecting first-order Markov processes. Sequential data may also be structured by higher-order processes, reflecting multi-lag effects that are not carried through first-order associations. The extension of EBREM to study higher-order processes remains to be worked out.

There is only limited information at this point on the data demands of this approach. Since power depends in a complex way on the number of episodes and observations per episode, number of observational codes, and other model-specific factors, currently there are no simple ways to determine power and sample size for these models. The example we have used here involves an unusually large number of episodes or couples, when compared to other recent reports of sequential data. Rather than provide extensive power analyses at this point, instead we tested whether a study involving only half the number of couples would find significant effects as well. We analysed the model of stationarity in the presence of heterogeneity on a randomly selected 50% of couples. This half sample test reached 0.021 significance while the full sample reached 0.005 significance, suggesting that the effects are strong enough to be detected with a substantially smaller sample. More work on the statistical power of these methods is in order to help investigators understand the necessary sample sizes and episode lengths required to detect patterns of interest.

A final limitation of the empirical Bayes estimation approach is that it does not take into account the uncertainty in parameter estimates of the covariance components. Fully Bayesian approaches can take into account such uncertainties in variance components estimation. Recent advances in computation, such as that employed by BUGS software (Spiegelhalter, Thomas, Best, & Gilks, 1995), may make this approach practical, since these programs are flexible enough to incorporate the forms of hierarchical structure discussed here. It still needs to be determined whether and under what conditions this source of variation could compromise interpretation of findings based on the empirical estimation approach.

References

Allison, P. D. & Liker, J.K. (1982). Analyzing sequential categorical data on dyadic interaction. *Psychological Bulletin*, 91, 393-403.

Bakeman, R. & Gottman, J. M. (1986). *Observing interaction: An introduction to sequential analysis.* Cambridge: Cambridge University Press.

Bakeman, R. & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis* (2nd ed.). Cambridge: Cambridge University Press.

Bakeman, R. & Quera, V. (1995). Log-linear approaches to lag-sequential analysis when consecutive codes may and cannot repeat. *Psychological Bulletin*, 118, 272-284.

Bank, L., Petterson, G. R., & Reid, J. B. (1996). Negative sibling interaction patterns as predictors of later adjustment problems in adolescent and young adult males. In G.H. Brody (Ed.), *Sibling relationships: Their causes and consequences. advances in applied developmental psychology* (pp. 197-229). Norwood, NJ: Ablex Publishing Corp.

Brown, C. H. & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiology paradigm. *American Journal of Community Psychology*, 27, 673-710.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Carlin, B. P. & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis.* New York: Chapman and Hall.

Davis, B. T., Hops, H., Alpert, A., & Sheeber (1998). Child responses to parental conflict and their effect on adjustment: a study of triadic relations. *Journal of Family Psychology*, 12, 163-177.

DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.

Eckerman, C. O. (1993). Toddlers′ achievement of coordinated action with conspecifics: A dynamic systems perspective. In L.B. Smith & E. Thelen (Eds.), *A dynamic systems approach to development: Applications* (pp. 333-357). Cambridge, MA: MIT Press.

Gardner, W. (1990). Analyzing sequential categorical data: Individual variation in Markov chains. *Psychometrika*, 55, 263-275.

Gottman, J. M. (1979). *Marital interaction: Experimental investigations.* New York: Academic Press.

Gottman, J. M. (1980). Analyzing for sequential connection and assessing interobserver reliability for the sequential analysis of observational data. *Behavioral Assessment*, 2, 361-368.

Gottman, J. M. & Roy, A. K. (1990). *Sequential analysis: A guide for behavioral researchers.* Cambridge: Cambridge University Press.

Hedeker, D & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, 50, 933-944.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando: Academic Press.

Howe, G. W. (1995). *Couples problem solving following job loss.* Paper presented at the Biennial Meeting of the Society for Research in Child Development, Indianapolis, IA.

Kiser, L. J., Bates, J. E., Maslin, C. A., & Bayles, K. (1986). Mother-infant play at six months as a predictor of attachment security at thirteen months. *Journal of the American Academy of Child Psychiatry*, 25, 68-75.

Margolin, G. & Wampold, B. E. (1981). Sequential analysis of conflict and accord in distressed and nondistressed marital partners. *Journal of Consulting and Clinical Psychology*, 49, 554-567.

Minuchin, S., Rosman, B. L., & Baker, L. (1978). *Psychosomatic families: Anorexia nervosa in context* (pp. 39-45). Cambridge, MA: Harvard University Press.

Muthén, L. K. and Muthén, B. O. (1998). *Mplus User's Guide.* Los Angeles, CA: Muthén & Muthén.

Patterson, G. R. (1979). A performance theory of coercive family interaction. In R.B. Cairns (Ed.), *The analysis of social interactions: Methods, issues, and illustrations* (pp. 119-162). Hillsdale, NJ: Lawrence Erlbaum.

Revenstorf, D., Hahlweg, K., Schindler, L., & Vogel, B. (1984). Interaction analysis of marital conflict. In K. Hahlweg & N.S. Jacobson (Eds.). *Marital interaction: Analysis and modification* (pp. 159-181). New York: The Guilford Press.

Sackett, G. P. (1979). The lag sequential analysis of contingency and cyclicity in behavioral interaction research. In J.D. Osofsky (Ed.), *Handbook of infant development* (pp. 623-649). New York: Wiley.

Snyder, J., Edwards, P., McGraw, K., & Kilgore, K. (1994). Escalation and reinforcement in mother-child conflict: Social processes associated with the development of physical aggression. *Development and Psychopathology*, 6, 305-321.

Sörbom, D. (1989). Model modification. *Psychometrika,* 54, 371-384.

Spanier, G. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family*, 38, 15-28.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. R. (1995). *BUGS: Bayesian inference using Gibbs sampling, Version 0.603.* Medical Research Council Biostatistics Units: Cambridge.

Takkouche, B., Cadarso-Suárez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, 150, 206-215.

Upton, G. J. G. (1982). A comparison of alternative tests for a 2 × 2 comparative trial. *Journal of the Royal Statistical Society*, Series A [General], 145, 86-105.

Wickens, T. D. (1993). Analysis of contingency tables with between-subjects variability. *Psychological Bulletin*, 113, 191-204.

Appendix A

Splus program for computing iteratively the maximum likelihood estimates of mu and tau.sq based on Equations (6) and (7).

```
    # − − − − − − − − − − − − − − − − − − − − − − − − − − − − − − −
#Data:
# odds −− > contains the log odds ratios
# odds.var −− > contains variances of log odds ratios
# L −− > the number of units in the study (e.g., couple)
#− − − − − − − − − − − − − − − − − − − − − − − − − − − − − − −−
# Initial values
iter < − 0
del < − 2
mu < − 1
tau.sq < − 0
while(abs(del) > 1.0e − 10  &&  (iter < −iter + 1) < 100) {
del < −mu
wlam < −1/(odds.var + tau.sq)
mu < −sum(wlam ∗ odds)/sum(wlam)
del < −mu − del
tau.sq < −sum(wlam∗wlam∗((odds−xmu)∗(odds−xmu)−odds.var))/sum(wlam∗wlam)
}
write( mu, tau.sq)
```

Appendix B

Alternative Specification for Nonstationarity

We propose an alternative specification of a random effects model for testing stationarity between two time periods for each episode.

Level 1:

$$\begin{aligned} Y_{m_1} &= \theta_{m_1} + E_{m_1} \\ Y_{m_2} &= \theta_{m_2} + E_{m_2} \end{aligned}$$

Level 2:

$$\begin{aligned} \theta_{m_1} &= \mu + V_{m_1} \\ \theta_{m_2} &= \alpha + \gamma\theta_{m_1} + V_{m_2} \end{aligned}$$

where, $V_{m_1} \sim N(0, \tau_1^2)$, $V_{m_2} \sim N(0, \tau_2^2)$ and V_{m_1} and V_{m_2} are uncorrelated but the covariance between θ_{m_1} and θ_{m_2} is $\gamma\tau_1^2$ indicating that the covariance is proportional to the variability in period 1.

Author's Note

Getachew A. Dagne is Assistant Professor, Department of Epidemiology and Biostatistics, University of South Florida, Tampa. George W. Howe is Professor, Department of Psychiatry and Behavioral Sciences, The George Washington University, C. Hendricks Brown is Professor, Department of Epidemiology and Biostatistics, University of South Florida, Tampa, Bengt O. Muthén is Professor, Graduate School of Education & Information Studies, University of California, Los Angeles. The authors thank the Editor, Associate Editor, and four anonymous reviewers for their careful reading and detailed comments that have greatly enhanced this article. We would also like to thank our methodological colleagues at the Prevention Science and Methodology Group (PSMG) who have provided many insights. Special acknowledgement is made to our colleague Mike Stoolmiller at the Oregon Social Learning Center for helpful comments and suggestions. Work on this paper was supported by the NIMH Prevention, Early Intervention and Epidemiology Research Branch, the Child and Adolescent Treatment and Preventive Intervention Research Branch, and the Adult Geriatric Treatment and Preventive Interventions Research Branch under Grants No. MH40859, "Designs and Analyses for Mental Health Preventive Trials," No. MH01259, "Methodologic Advances in Mental Illness Prevention," No. MH38725, "Epidemiologic Center for Early Risk Behaviors," and No. MH42968, "Periodic Outcome of Two Preventive Trials." This article is also based on research supported in part by funding from the Prevention Research Branch of the National Institute of Mental Health (R01-MH47292).

Correspondence can be sent to Getachew Dagne, Department of Epidemiology and Biostatistics, University of South Florida, 13201 Bruce B. Downs, MDC 56, Tampa, FL 33612. Electronic mail may be sent to: gdagne@hsc.usf.edu.

Footnotes

[1] We use this 4×4 table here only to illustrate some properties of adjusted cell residuals. As we note below, analyses of residuals based on the full table are usually not valid, since this combines different types of transitions, which are usually analyzed separately in smaller sub-tables, such as those used in the remainder of this paper.

[2] Many coding systems will include more than one coding category for relevant antecedents. For example, Margolin and Wampold (1981) included categories for positive, negative, and neutral behavior, forming 3×3 subtables for male to female transitions. One possible generalization to these situations would involve collapsing all relevant antecedent categories other than the antecedent of interest into one "other antecedents" category. We plan to address the complexities of studying larger contingency tables in future work.

[3] The level 2 model we advance here follows the logic of confirmatory factor analysis, and will be applicable in most cases where changes in stationarity are not extreme. However, it would lead to incorrect specifications in extreme cases such as when the contingency actually reverses sign over the course of the episode. We have developed an alternative specification where the covariance between the random effects is proportional to the variability in period 1, which may be useful in these situations. A brief account of this model can be found in Appendix B.

[4] Gardner (1990) has presented a method for studying covariate effects on entire contingency tables based on sequential data, but without checking for the presence of heterogeneity.

[5] An anonymous reviewer pointed out, quite rightly, that extra parameters may lead to instability in estimates, and that a simpler model ignoring nonstationarity, while containing some misspecification, may provide more stable estimates. Investigators can test whether ignoring nonstationarity has substantive effects on model parameters, or whether specifying the model without these parameters has minimal effect on other parameters while increasing the precision with which they are estimated.

Table 1

Raw Frequencies and Adjusted Residuals for a Single Observed Couple, and Compared to a Simulated Couple.

Table 1(a): Raw Frequency of Sequences for a Single Couple

	Consequent			
Antecedent	Male neg	Male pos	Female neg	Female pos
Male neg	49	7	30	27
Male pos	1	33	16	42
Female neg	27	12	21	8
Female pos	35	39	1	41

Table 1(b): Raw Frequency of Sequences for a Single Couple when Codes Cannot Repeat

	Consequent			
Antecedent	Male neg	Male pos	Female neg	Female pos
Male neg	0	7	30	27
Male pos	1	0	16	42
Female neg	27	12	0	8
Female pos	35	39	1	0

Table 1(c): Adjusted Residuals for a Single Couple

	Consequent			
Antecedent	Male neg	Male pos	Female neg	Female pos
Male neg	0	-4.29	5.74	-1.08
Male pos	-6.39	0	1.18	5.12
Female neg	4.70	-0.27	0	-4.23
Female pos	1.89	4.41	-6.62	0

Table 1(d): Adjusted Residuals for a Simulated Couple

	Consequent			
Antecedent	Male neg	Male pos	Female neg	Female pos
Male neg	0	-13.57	17.08	3.41
Male pos	-20.21	0	3.73	16.19
Female neg	14.86	0.85	0	-13.38
Female pos	5.98	13.94	-20.93	0

Table 2

Frequency of Two-step Sequences for mth Episodes and lth Unit

	Consequent (j)			
Antecedent(i)	1	2	$\cdots$	J
1	n_{11ml}	n_{12ml}	$\cdots$	n_{1Jml}
2	n_{21ml}	n_{22ml}	$\cdots$	n_{2Jml}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	n_{I1ml}	n_{I2ml}	$\cdots$	n_{IJml}

Table 3

Log Odds Ratio (Variance) for a 2×2 for a Single Couple

	Consequent	
Antecedent	Female Neg	Female Pos
Male Neg	1.07 (.16)	-1.07 (.16)
Male Pos	-1.07 (.16)	1.07 (.16)

Table 4

Model Fit Summary for Test of Stationarity and Heterogeneity

Table 4(a): Test of Stationarity in the Presence of Heterogeneity

Model*	$-2LL$	G^2	ΔDF	p-value
(1) unrestricted	740.58			
(2) $\mu_1 = \mu_2, \tau_1^2 = \tau_2^2 = 0$	753.45	12.97	3	.005
(3) $\mu_1 = \mu_2$	747.19	6.62	1	.010
(4) $\tau_1^2 = \tau_2^2 = 0$	745.60	5.02	2	.081

Table 4(b): Test of Heterogeneity in the Presence of Nonstationarity

Model	$-2LL$	G^2	ΔDF	p-value
(5) $\psi^2 = 0, \tau_1^2 = 0,\ \tau_2^2 = 0$	780.03	39.45	3	$< .001$
(6) $\psi^2 = 0$	746.35	5.77	1	.016

* All estimates of parameters subject to the specified restrictions.

Table 5

Estimates of Parameters from Unrestricted Model

Parameter	Value (95% CI)
μ_1	1.32 (1.18-1.47)
μ_2	1.58 (1.43-1.72)
τ_1^2	.18 (.00 - .45)
τ_2^2	.15 (.00 - .40)
ψ^2	.18 (.02 - .34)

Table 6

Summary of Model Fitting

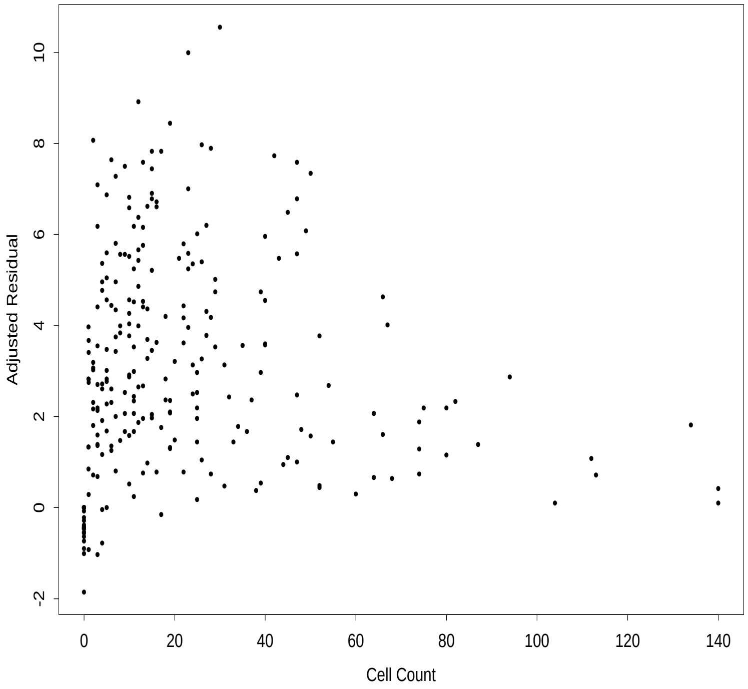
	Unrestricted		$\beta_{11} = \beta_{12}$	
Parameter	Estimate	SE	Estimate	SE
μ_1	.35	.474	.30	.342
β_{11}	.01	.004	.01	.003
μ_2	.51	.477	.57	.345
β_{12}	.01	.004	.01	.003
τ_1^2	.18	.148	.19	.148
τ_2^2	.17	.132	.17	.132
ψ^2	.13	.079	.13	.079
$-2LL$	695.64		696.41	

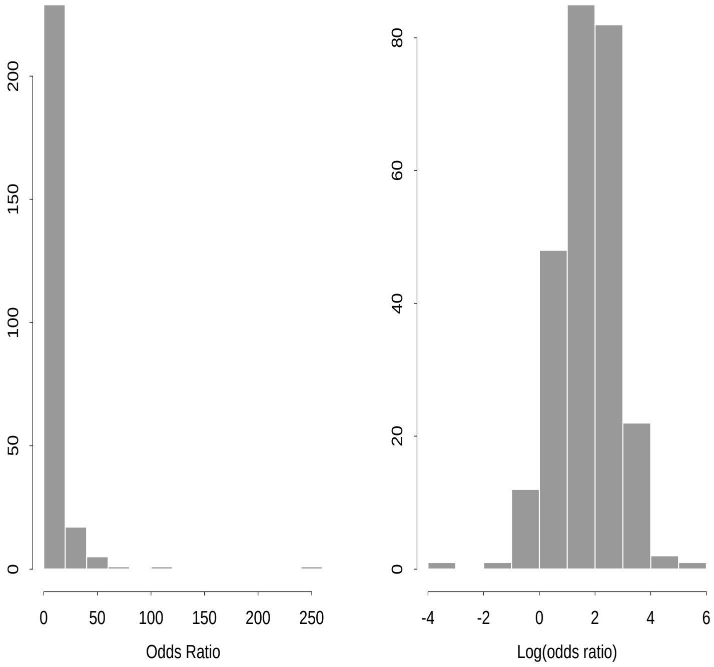
Figure Captions

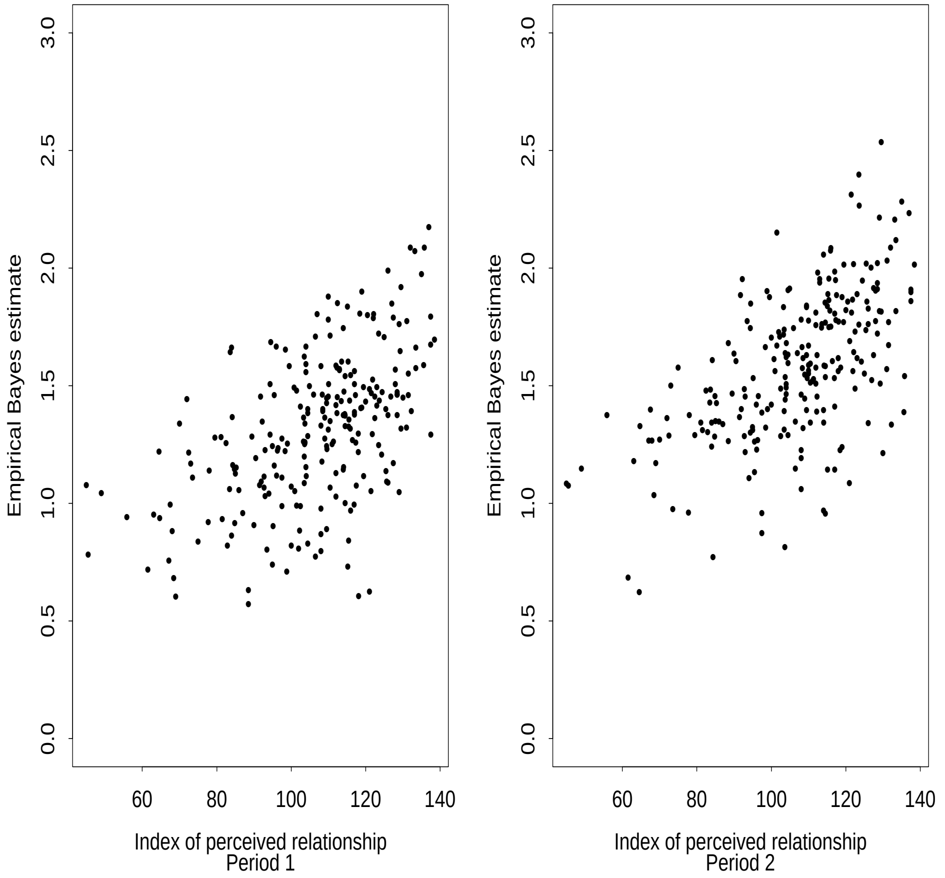
Figure 1. Plot of adjusted residuals and observed counts

Figure 2. Histograms of odds ratios and log odds ratios for male Neg Vs female Neg

Figure 3. Scatter plot of estimated log odds ratios and index of perceived relation


Adjusted Residual
-2
0
2
4
6
8
10
0
20
40
60
80
100
120
140
Cell Count


0
50
100
150
200
0
50
100
150
200
250
Odds Ratio
0
20
40
60
80
-4
-2
0
2
4
6
Log(odds ratio)


Empirical Bayes estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
60
80
100
120
140
Index of perceived relationship
Period 1
Empirical Bayes estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
60
80
100
120
140
Index of perceived relationship
Period 2