Introduction

 Multivariate, multilevel survival analysis (MMSA) is a statistical model that for our purposes estimates 2 or more correlated latent hazard rates from the observation of repeated events within subjects. The subject level hazard rates can be embedded in a larger SEM framework and used as outcomes or predictors, or both, the last two uses being very recent advances that makes MMSA potentially a very powerful statistical tool for researchers interested in coercive processes. Using subject level hazard rates as predictors, we focus on the application of simple MMSA models to emotional behaviors of children and their parents on two occasions during structured problem solving tasks in the lab and how well the rates are both predicted by concurrent child antisocial behavior and predictive of change in future child antisocial behavior. MMSA is currently available in at least one commercial software package, the Mplus program (Muthen & Muthen, 2013) and we provide links to Mplus syntax for fitting all of the MMSA models.

 Our presentation is organized in the following manner. First we introduce terminology and discuss a few basics about MMSA like hazard rates and censoring in a non-technical way to give readers a brief overview. Then we relate hazard rates to other common measures derived from direct observation of behavior and discuss the advantages and disadvantages of taking a multilevel approach. We introduce our sample and then walk readers through a series of models designed to illustrate the application of MMSA. We also offer practical advice for moving beyond basic models. Finally we discuss the current computational limitations of MMSA and identify issues in need of further methodological development.

 To start at the beginning, the hazard rate is fundamental to MMSA and is the rate at which events happen in the moment given that a subject is at risk, that is, capable of experiencing the event. The simplest example and the one that has driven most of the development of the method and hence the terminology is the event of death. To be at risk of death, you must currently be alive. The hazard rate gives the death rate. The higher the hazard of death, the more likely a currently living subject will die in the next short period of time. For behavioral observation, the hazard rate usually refers to repeatable, everyday behaviors, not so called trapping states like death. As such, it represents the likelihood of a subject performing the target behavior given that they are currently capable of doing so which at a minimum means they are not currently engaging in the behavior. You are not at risk of performing the behavior if you are currently performing the behavior.

 If we are recording behaviors because we are interested in models for hazard rates, we have to deal with the situation where the target behavior does not happen during the time we have available to observe. We call this simple right censoring (SRC). Models for hazard rates incorporate information about SRC to avoid biased estimates. SRC always happens unless we can observe subjects continually and forever. Other types of censoring commonly happen in MMSA as well, the most important being so called “competing risks censoring” (CRC).

 To understand CRC, suppose a parent can be in one of 3 states, negative (A), neutral (N), or positive (P) during an observation session. If the parent is currently neutral, N, then the parent is at risk for a state change to A or P. The state with the higher hazard rate will tend to happen first (say for example, P) and this event will censor the observation of a state change to A. P and A are competing and the winning state P has its winning time recorded along with the fact that it won and the losing state A has the same time recorded but along with the fact that it lost, i.e., was censored by P. CRC can be extensive in observation data and must be incorporated in MMSA to avoid biased estimates of hazard rates.

 Figure 1 shows a heuristic path diagram for a simple MMSA with 2 hazard rates, the initiation rate of child anger for session 1 and 2. These rates are predicted by concurrent child antisocial behavior (fall K child antisocial) and in turn predict change in future child antisocial behavior in the spring of grade 3. We will use the model in Figure 1 extensively to illustrate various applications of MMSA including models for estimating retest reliability of a single occasion (i.e., session) of observation and models for change across two sessions of observation. The diagram is divided in two by the heavy black line. Above the line is the subject level model and below is the within subject level model. The within subject level model is made up of repeated durations of states of interest or episodes, in this case AC is shorthand for child anger, NotAC is shorthand for child not anger (all other possible child states) and the arrow between NotAC and AC indicates that the duration of NotAC ended in AC (i.e., each episode is how long it took the child to get angry). The within subject model also includes a censoring indicator that indicates whether the NotAC duration was censored or not (yes = 1, no = 0). The duration variable and the censoring indicator are shown in the same box to indicate that they must go together in order to estimate the model but the dividing line indicates that they are two separate pieces of information. When we collapse all behavior down to two alternating states, the only censoring is SRC, it is not usually extensive and it happens once when the session ends for each subject.

 Most of the subject level model follows the usual conventions for path diagrams of SEM’s. The wiggly arrows, however, indicate that the mathematical relation between observed durations and the latent hazard is nonlinear and in fact, the latent hazards, f1Ac and f2Ac are on the log hazard scale. Also, playing a role similar to that of a regression intercept is the log of the baseline hazard function for each latent hazard rate. In fact, models in this paper will be very simple and employ a constant baseline hazard function so that the parallel with regression intercepts is accurate. This is not always the case with MMSA and the MMSA can accommodate parametric (e.g., Weibull-- J or upside down J shaped) or non-parametric (e.g., Cox) representations for non-constant baseline hazard functions. Human lifespan is a familiar illustration of a non-constant baseline hazard function. While “waitin’ round to die”, (i.e., at risk of death) the death rate is changing starting out higher at birth, declining and flattening out after we survive infancy and then gradually increasing again as we head in to our golden years.

 Although it is beyond the scope of this paper, MMSA can be extended to the situation where the hazard rates for behavior of one member of the dyad are conditional on the behavior of the other member of the dyad (e.g., the hazard of child anger given parent negative, see Stoolmiller & Snyder, 2013). This is a very brief conceptual introduction to MMSA and more information is available on the author’s ResearchGate home page and in other sources (Stoolmiller & Snyder, 2006).

 Before we start fitting MMSA models, it is valuable to put hazard rates in perspective with other traditional measures derived from observation. This is relatively straightforward if we stick to the simple single level survival model (data for one subject only) with a constant hazard rate. In this case, the best estimate (i.e., maximum likelihood estimate) of the hazard rate for a single subject is the total number of times the subject visited the target state divided by the total time they were at risk for such a transition (Gardner, 1993). For example, in 60 minutes of observation, suppose a child visited the target state of neutral 50 times and spent a total of 20 minutes in neutral. The total risk time is therefore 40 minutes because the child is not at risk of a transition to neutral while they are neutral and the hazard rate is 50/40 = 1.25 events per minute or 75 events per hour of risk. If the child went to neutral 50 times, then they left neutral about 50 times (depending on what state they started in but the minor discrepancy is not particularly important) which means the hazard of neutral termination is 50/20 = 2.5 events per minute or 150 events per hour of risk. Given these two hazard rates, we can determine average durations, duration proportions (prevalence), simple rates (incidence) and frequency proportions.

 The incidence rate for child neutral is 50/60 = 0.83 events per minute or 50 events per hour but the incidence rate is determined by the hazard rates for neutral initiation (i.e., going to neutral) and neutral termination (i.e., leaving neutral). The incidence rate is the product of the hazards divided by the sum of the hazards which is (50/40 \* 50/20)/(50/40 + 50/20) = 5/6. The average duration for child neutral is 20 minutes divided by 50 visits or .4 minutes per visit but it is also the reciprocal of the hazard of neutral termination, 20/50 = .4 minutes per visit. A similar relation holds for the average duration of not-neutral. Neutral duration proportion (prevalence or conditional probability) is the proportion of total time spent in neutral and is 20/60 = .33 but this is also determined by the ratio of the hazard of neutral initiation to the sum of the hazards for neutral initiation and termination (50/40)/(50/40 + 50/20) = 1/3. A similar relation holds for not-neutral duration proportion. Neutral frequency proportion is the proportion of events that were neutral and is 50/100 = .5 but this is also determined by the ratio of the incidence rate for neutral to the total incidence rate for all states, which in turn is determined by the hazard rates.

 The important point is that fitting models for hazard rates does not leave out other important aspects of interaction such as average duration, prevalence and incidence rates. These quantities are nonlinear functions of the hazard rates and may be of more or less direct interest depending on the application but they are completely determined by the hazard rates. This is also true for measures which have popped up in the literature in studies taking a dynamical systems perspective on behavioral regulation, measures derived from average duration, prevalence and incidence rates. For example, total transitions (e.g., Hollenstein, Granic, Stoolmiller & Snyder, 2004) is just the sum of all the hazard rates, dispersion (e.g., Lunkenheimer et al., 2011) is a standardized version of the sum of the squared duration proportions and entropy (e.g., Dishion, Forgatch, Van Ryzin, & Winter, 2012) is the negative sum of the products of the duration proportions times their respective logs. These last two essentially measure the same thing and are both nonlinear transformations of prevalence.

 It is also valuable to put the multilevel aspect of MMSA in perspective. A fair question is that if the hazard rate for a single subject is so easily computed (as demonstrated above), why is it necessary to build a complicated multilevel model to estimate these as latent rates just because we have a sample of subjects? There are two sets of advantages and the first set is more straightforward so we start there. If we are interested in within subject level research questions, then we need a multilevel model. There are many potentially interesting within subject questions because these sorts of questions get at what shapes moment to moment behavior which is of obvious interest to researchers of coercive family process (see for example Dagne & Snyder, 2009; Dagne & Snyder, 2011; Snyder, Stoolmiller, Wilson & Yamamoto, 2003).

 If we do not have within subject questions, however, the advantages of parameter estimates from a latent multilevel model (hereafter referred to as LML estimates) over a collection of simple single level estimates for the same sample (the procedure proposed by Gardner, 1993, hereafter referred to as CSL estimates) are not so obvious. Given the complexity of the multilevel model and the current limitations due to computational intensity, there are tradeoffs involved and although we do not have all the answers, we revisit some of the issues raised by Gardner (1993) to help readers make more informed choices.

 The computational intensity comes because the nonlinear aspects of the hazard model make it necessary to use a technique known as numerical integration (in multiple dimensions, usually one dimension for each hazard rate) in an iterative fashion to search for the LML estimates. The computational intensity grows exponentially with each additional dimension of integration required, and the method is quite sensitive to the initial values we use to start the procedure searching. Even if we guess at very good starting values (we give some advice on this later), on currently available desk top computers, it can take many hours and even days for the procedure to finish. Given this, we cannot afford to fall in love with MMSA just for the sake of having a fancy modeling technique, MMSA needs to provide some very real advantages over the CSL approach.

 In broad terms, MMSA has the same advantages that multilevel models for normally distributed continuous scale scores (e.g., growth curve models or random school regression models) have over CSL approaches to the same data (see Raudenbush & Bryk, 2002, Ch. 1). One could argue that since multilevel models have virtually eliminated the CSL approach in this domain, they must be vastly superior but this may not always be the case because computational time is rarely an issue. In other words, multilevel models for normal outcomes run so fast that there is very little cost to switching to a multilevel approach, even very complex models can be entertained, and in fact, in most cases theoretical complexity is struggling to catch up with the multilevel methodology. With MMSA, the reverse seems more accurate, there is a high cost of switching to MMSA, computational intensity limits MMSA to relatively simple models and the multilevel methodology is still struggling to catch up with theoretical complexity.

 Conceptually, the multilevel latent log hazard for an individual subject at a single session can be thought of as an optimally weighted combination of their single level log hazard (ML estimate computed for a single subject in isolation) and the population level log hazard (ML estimate computed in the MMSA) from which the subject was sampled, where the weights are determined by the precision of the single level estimate. The more precise the single level estimate, the greater the weight it receives and the less precise the single level estimate, the more weight the population level estimate is given. In fact, it is possible to compute the reliability of the CSL estimates as proxies for the latent rates in MMSA. Clearly, if the reliability was low, we should stick with the MMSA but if the reliability was high, then there would be little difference and we could use either the CSL or the LML approach and get about the same answer. Except, nobody does this calculation to determine which model to use for normal outcomes because there is so little cost to using the multilevel model; it just gets used. At the current state of development of MMSA, however, it might be worth using a CSL approach at times as advocated by Gardner 20 years ago or perhaps in combination with simpler MMSA models.

 Thus, it is worth thinking about the major influences on the precision of the single level estimate of a given hazard rate. One major influence is the amount of time we observe the subject also known as variance due to finite time sampling. For models with constant hazards, the longer we observe a subject, the more events happen and the more precise our estimate of the subject’s hazard rate becomes. This also means that for a fixed amount of observation time, we will obtain more precise estimates of higher hazard rates than lower hazard rates for a given subject. Thus, the reliability of CSL estimates is strongly determined by the amount of observation time and the true level of the hazard rates. We will refer to this type of reliability as finite time sampling (FTS) reliability. Coder errors will also impact CSL estimates but unless we have 2 or more coders coding the same subject at the same session, this source of variance will be invisible to both CSL and MMSA. Rogosa and Ghandour (1991) argued that FTS reliability is substantial and in their simulations it was usually the more important source of error given reasonably well trained coders. For this reason, we will not pay much attention to coder reliability in this paper. We will, however, come back to this issue in future work.

 If we consider two sessions of observation using a different coder for each subject at each session, then we have retest reliability. This is the bottom line for us because the level of retest reliability will be determined by sources of error such as finite time sampling at each session, coder errors and session specific influences on the subjects, and individual differences in true hazard rates at a particular point in developmental time (e.g., fall of Kindergarten). Retest reliability tells us how much signal we have (variance in true hazard rates) compared to signal plus noise from all sources. The relations of true hazard rates with other variables, free from contaminating noise is what we seek to study either via MMSA or CSL.

 If we only have a single session of observation, we are at a distinct disadvantage in that we cannot determine the retest reliability. We can and should still deal with FTS reliability. Paradoxically, in order to tell when Gardner’s CSL approach is good enough and when it might be better to switch to MMSA, we have to actually fit an MMSA. The good news, however, is that a simple MMSA that runs quickly is probably adequate to make this decision, which we demonstrate below. Eventually, as computing speed continues to increase (due to improvements in hardware, software, parallel processing, etc.), these issues will go away for MMSA just as they did for multilevel models for normal outcomes.

 Having introduced the basic model, we next introduce the sample and the measures, and then turn to a series of research questions that illustrate the model, provide some insights in to the role of emotion regulation in the development of antisocial behavior, and investigate how parent-child dyads change their behavior in response to the demand characteristics of a typical problem solving observational paradigm.

Methods

 The sample, measures and lab paradigm have been described extensively in previously published work so only a brief description is provided here. The participants were 275 children, (and their parents), 49% male, who were recruited at kindergarten entry to one elementary school that exclusively served a low-income, mixed industrial-residential neighborhood. Seventy-one percent of the children were European Americans, 19% were African-American, 5% were Hispanic/Latino, and the remaining children were Native- and Asian-American. The recruitment rate of children at kindergarten entry was 76%.

 Child and Parent Emotion Displays. The interaction of parent-child dyads was videotaped for 2 hours on each of two occasions, separated by a minimum of 1 week (average = 2.3 weeks). The structured activities of the 2 sessions were very similar but not identical. The first hour of interaction was coded using the Specific Affect Coding System (SPAFF; Gottman, McCoy, Coan, & Collier, 1996). Observers were trained extensively to a criterion of 75% agreement prior to initiating coding of videotaped interaction derived from participants in the current research. Weekly recalibration training was completed to minimize observer drift. The average between-observer agreement on the occurrence of SPAFF codes (using a +/-6 second window) was 83% (kappa = .73).

 In the hazard analyses, super-ordinate categories labeled child “anger”, ”positive”, “sad-fearful”, and “neutral”, were defined by combining the SPAFF codes of anger, contempt, and disgust, SPAFF codes of validation, interest, enthusiasm and humor, SPAFF codes of sad and fear-tension and all other codes respectively. Super-ordinate categories labeled parent “negative”, ”positive”, and “neutral”, were defined by combining the SPAFF codes of anger, contempt, disgust, criticism, domineering, belligerence, threats, whining, stonewalling, defensiveness, fear/tension, and sadness, SPAFF codes of validation, interest, enthusiasm and humor, and all other codes respectively. For both parent and child, coders could also use an out-of-view code if either subject was off camera and their emotional state could not be accurately coded. Preliminary analyses indicated that hazard rates for child sad/fear and out-of-view for both parent and child did not change significantly across the sessions and were not related to child antisocial behavior so they are not included in any of the analyses in this paper.

 Child Antisocial Behavior. A measure of child antisocial behavior was obtained in the fall (child age 5.5 years) of the kindergarten year and in the spring of 3rd grade consisting of the mean of 23 items scored 0, 1 or 2 from the Teacher Rating Form (Achenbach & Edelbrock, 1991) describing aggressive and oppositional child behavior. The scale reliabilities (alpha) were greater than .90 on each measurement occasion. See Snyder et. al., 2003, for a more thorough description of the sample, the family level measures, the observational procedures and coding for parent-child emotion displays.

Results

Research Question 1: Assuming a constant baseline hazard function, do the average hazards for the child and parent states change from session 1 to 2?

 Table 1 shows descriptive statistics for the CSL hazard rates (rate per hour at risk) computed as described previously. To be comparable to MMSA results, the CSL hazard rates must be log transformed but a problem arises because for two of the hazard rates, child anger and parent negative, we have subjects that were completely censored and had no transitions in to those states by the end of a session. The log transformation is not defined for a value of zero so to make some progress, we use a mathematically reasonable suggestion attributed to Sorenson (1978) by Tuma and Hannan (1984) and give these subjects ½ of a transition first before dividing by their time at risk and then use the log transformation. This illustrates an advantage of MMSA because these subjects are incorporated in the MMSA without difficulty. Descriptive statistics for the log CSL rates and change in log CSL rates are also shown in Table 1.

 What is clear in Table 1 is that several of the CSL hazard rates changed significantly between session 1 and 2 including child anger and parent neutral, which increased, and parent positive which decreased substantially (t = -9.33, df =219, p < .001) by just over half a session 1 standard deviation (d = , d=-0.56). Parent anger increased slightly but the change was only marginally significant (p = .054). It would appear that the second session was less fun than the first. In fact, 19 dyads that participated in the first session declined to participate in the second. A reasonable hypothesis would be that if the parent found the first session to be unpleasant, they would be more likely to skip the second. A logistic regression analysis using the log CSL hazard rates reveals that parents with higher negativity in the first session (e.g., at the 90th percentile) had 5.7 times the odds or equivalently, 1.5 times the probability of skipping the second session compared to a parent with a low score (e.g., at the 10th percentile). None of the other parent or child hazard rates were predictive of skipping session 2.

 We can address the same questions using MMSA, sort of. Typically, with correlated data like the parent and child hazards, it would be most efficient to analyze all 6 hazards at both sessions simultaneously. This would result in 12 hazard rates and 12 dimensions of numerical integration which would completely overwhelm most desktop computers. So instead, we do 6 separate models, one model for each hazard at both sessions as in Figure 1, which only involves 2 dimensions of integration. In addition, to speed up calculations, we make sure that the data are structured so that only those rows of data with the durations and censoring indicators of interest are included in the model. For example the child anger model involves 2601 rows of data to represent the repeated durations and censoring indicators for both sessions. While it would be possible to use one large child file with all child transitions among all possible states for both sessions, the file would have 127790 rows of data and this would substantially slow down all the child models including the model for child anger which only needs 2601 rows. Finally, we use the log CSL hazard rates to provide reasonable starting values for the corresponding MMSA, which not only speeds up calculations but also enhances our chances of successful model convergence.

 Once we run all 6 MMSA models, we can also use the estimated standard errors of the empirical Bayes estimates of the latent log hazards to compute the FTS reliabilities of the CSL estimates as proxies for the latent hazards. If the FTS reliabilities are high, like .90 or higher, then we should get about the same results from either the CSL estimates or MMSA. This in turn would mean that we can eliminate some of the latent hazards and replace them with observed CSL estimates, reduce the dimensions of integration and computational intensity and entertain more complex models. This brings us to our second set of research questions.

Research Questions 2: How much of the variance of each hazard rate at each session is measurement error due to finite time sampling and what is the retest reliability of the hazard rates across both sessions?

 Table 2 shows the results for the 6 MMSA models with respect to mean hazards, FTS reliabilities at session 1 and 2 and retest reliabilities. As is clear, the MMSA results corroborate the results of the CSL analysis on mean hazards; child anger and parent neutral increased and parent positive decreased. In addition, the FTS reliabilities range from .80 to .98. The lowest reliabilities were obtained for child anger at both sessions. All other reliabilities were .90 or higher, which suggests that further analyses could safely use the corresponding CSL estimates.

 A separate MMSA also shows that the latent log hazard rate for parent negative significantly predicted skipping session 2 further confirming CSL results. In the latent analysis a parent high (e.g., 90th percentile) on the log hazard of negative had 5.5 times the odds or equivalently, 1.4 times the probability of skipping session 2 compared to a parent with a low score (e.g., 10th percentile). Since the FTS reliability of parent negative at session 1 was .90, it is really not surprising that these MMSA results are quite similar to the CSL results.

 Before we discuss results for the retest reliabilities, we revisit and expand on some of the issues discussed in the introduction. The significant shifts in 3 of the 6 mean hazards are interesting because all previously published hazard analyses of this data have combined the sessions. We did this largely out of practical necessity because it cuts the number of latent hazard rates in half and dramatically reduces the computational burden. Given the evidence of significant change it is worth considering two different types of models that explicitly account for change across the sessions.

 The first and simpler model is shown in Figure 2 and is sometimes called a random intercept model although we will refer to it as the reliability model. It adds a single higher order factor to the model in Figure 1, it can accommodate a change in the mean hazard and it can also accommodate individual differences in change between the sessions but assumes that the individual differences in change are due to time specific influences that are uncorrelated with all other variables in the model. Time specific influences include by definition random measurement errors including variation due to finite time sampling. As mentioned previously, because we used different coders for dyads across the two sessions, coder errors are also time specific influences. The higher order factor captures stable individual differences across the two sessions and allows us to use these individual differences as either outcomes, predictors or both. Models with hazards obtained by combining both sessions implicitly assume that the retest correlation is not significantly different from 1 and the reliability model can accommodate this as well but if the retest correlation is significantly and substantially lower than 1, effects from the combined sessions model will be seriously attenuated by the time specific influences reflected by the low retest correlations. The reliability model, however, by estimating the actual retest correlation corrects for the attenuation.

 The second alternative model is shown in Figure 3. It adds a second higher order factor to the random intercept model to capture individual differences in change and is usually known as a latent difference score model although we will refer to it as the change model. The change model is a simple growth curve model for two points in time designs but unfortunately, it has some important limitations. In order to identify and estimate the model, the variances of the residual time specific influences of the random intercept model must be set to fixed values. Zero is a possible fixed value but this is equivalent to saying that the variables have perfect reliability. In our case, a better choice would be a reasonable estimate based on coder reliability, either from some other source of information or an educated guess. Fixing the time specific variances makes it possible to have two correlated higher order factors that represent individual differences in change and intercept, where intercept can be conveniently defined to be initial status, final status or average level. The reliability model, however, and the change model are not nested for two points in time designs and in fact, the most general version of the reliability model has the same number of free parameters as the change model. Trying to pick between the two models can be difficult for two points in time designs.

 Because we have additional subject level variables in the model, however, as predictors and outcomes, (e.g., child antisocial behavior), we do have some additional information that could help pick between the models. If the reliability model is the correct model, then the latent difference score factor is really just composed of variation due to time specific causes and hence should not be predictable by other covariates and should not be predictive of other outcomes. On the other hand, if there really was systematic change from session 1 to 2, and if it is related to the child antisocial variables, then finding significant relations between the child antisocial variables and the change score would favor the change model. This suggests a model fitting strategy for each hazard rate such that we start by fitting the change score model and retain it if the change score factor is significantly related to child antisocial. If not, then we eliminate the change score factor and use the simpler reliability model. If in addition, we find that the retest correlation is not significantly different than 1 in the reliability model, then we eliminate the separate session hazards and further simplify the model to a single hazard for both sessions, perhaps with adjustments for change in mean hazards.

 Table 2 shows the MMSA retest correlations between the latent hazards at sessions 1 and 2 and between change in the hazards and fall K and 3rd grade child antisocial. All of the retest correlations were significantly and substantially different from 1.0, ranging from .65 down to as low as .16. The only significant correlation between change in log hazards and the child antisocial variables was for change in child log neutral which was positively associated with 3rd grade child antisocial (r = .19, z = 2.56, p = 0.0105). The significant correlation suggests that at least some of the change from session 1 to 2 was not due to time specific influences and the change model is a better choice than the reliability model for child neutral. To follow this up, we fit the MMSA version of the reliability model in Figure 2 for child neutral. The results reveal that the shared variance in child neutral across the two sessions is not significantly related to either fall K antisocial or 3rd grade antisocial.

 At this point, the evidence we have assembled before us indicates that 1) with the possible exception of child anger the simple CSL log hazard rates are adequate estimates of the latent log hazards, 2) all the retest correlations for these hazards are significantly and substantially lower than 1 so the reliability model is necessary for each hazard to avoid serious attenuation of effect sizes and 3) child neutral should also use the change model to study correlations with child antisocial. Putting all of this together, we estimate 2 versions of the full model with all hazards simultaneously. First, we use the CSL estimates and the reliability model for all the hazard rates. Second, we use the CSL estimates and the reliability model for all the hazards except for child neutral, for which we use the change model. Finally we check the 2 models above by re-estimating them using MMSA but just for child anger. In each version of the model we include the fall K child antisocial predictor and the 3rd grade child antisocial outcome. Our goal is to determine which if any of the hazards are correlated with the antisocial variables and which if any are uniquely predictive of changes in future 3rd grade child antisocial controlling for K child antisocial.

 Results for the reliability version of the combined hazards model are shown in Table 3. The first version of the model had an adequate fit but was substantially improved by adding covariances between time specific influences for child and parent positive, one at each session and child and parent neutral, one at each session. The nested chi-square improvement for adding these 4 parameters was substantial, 109.1 with 4 df, p < .0001. These parameters are interesting in their own right because they may represent a within session reciprocity process that is quite strong and goes well beyond the level of the more trait like correlations represented in the higher order factors but we will not pursue these effects here. The robust chi-square for the final version of the model was 73.65, df = 59, p = .0949, RMSEA=.030 [90% CI = .000, .050], indicating a reasonably good fit to the data. Fall K child antisocial was significantly correlated with 3 of the 6 log hazards, parent negative (r = .24, z = 3.04, p = .0024), parent neutral (r = -.27, z = -1.98, p = .0479) and child positive (r = -.39, z = -4.64, p < .001). Of the 6 log hazards, 3 were also significantly correlated with 3rd grade antisocial, parent negative (r = .23, z = 2.35, p = .0187), parent positive (r = -0.25, z = -3.04, p = .0024) and child positive (r = -.19, z = 2.16, p = .0310).

 In the regression equation with all 6 log hazards simultaneously predicting 3rd grade antisocial, none were significant unique predictors net of fall K child antisocial despite the significant bivariate associations for 3 of the 6 log rates. To understand this result, we followed up by estimating a separate MMSA model for each log hazard using each one as a predictor of 3rd grade antisocial net of fall K antisocial. Only parent positive had a significant effect which indicates that the significant correlations for parent negative and child positive with 3rd grade antisocial are due to confounding with fall K antisocial. The lack of effect for parent positive in the simultaneous equation with all 6 other hazards was due to its positive (but non-significant) correlation with parent neutral (r = .239, z = 1.133, p = .257). When parent neutral was removed from the equation, parent positive had a significant effect on 3rd grade antisocial net of fall K antisocial. Parent neutral, however, was not a significant predictor of 3rd grade antisocial when parent positive was removed from the equation. Thus, with some additional investigation the reliability model identifies only one significant unique predictor of 3rd grade child antisocial, parent positive, with a standardized effect of -.258 (z = -1.993, p = .046).

 To estimate the change version for the combined hazards model, we converted both parent and child neutral to change models to keep the constructs parallel to each other and accommodate their correlations with each other. The reliability model was maintained for all other hazards. We included the same covariances among the time specific influences for child and parent positive at sessions 1 and 2 as in the reliability model. For the parent and child neutral change models, to make sure that findings were not due to an over-optimistic assumption of perfect coder agreement, we fixed the time specific variances of the parent and child neutral log hazards at 20% of the total variance at session 1 as reasonable guesses at the level of coder reliability (80%). Results are shown in Table 4 and are very similar to results for the reliability model. The pattern of significant correlations with fall K and 3rd grade antisocial is the same as in the previous model except that change in child neutral is significantly correlated with 3rd grade child antisocial (r = .19, z = 2.44, p = .0146). No other change correlations were significant.

 Similar to results for the reliability model, in the regression equation with all 8 log hazards simultaneously predicting 3rd grade antisocial, none were significant unique predictors net of fall K child antisocial. As before, we followed up by estimating separate MMSA change models, one for parent and one for child neutral hazards using each one as the sole predictor of 3rd grade antisocial net of fall K antisocial. Only change in child neutral had a significant effect. Similar to the reliability model, the effect of parent positive was masked by parent neutral. When parent neutral and change in parent neutral were removed from the equation, parent positive had a significant effect. Neither parent neutral nor change in parent neutral, however, were significant predictors of 3rd grade antisocial when parent positive was removed from the equation. The effect of change in child neutral was masked by both child positive and child negative. When these two were removed from the equation, change in child neutral had a significant effect. Thus, with some additional investigation, the change model identifies two significant unique predictors of 3rd grade child antisocial, parent positive, with a standardized effect of -.155 (z = -2.026, p = .043) and change in child neutral, with a standardized effect of .161 (z = 1.973, p = .049).

Discussion

 The purpose of this paper was to demonstrate the strengths and weaknesses of using MMSA to investigate stability and short term change in hazard rates of emotion for parents and children across 2 sessions of observation, when change was not really expected, and how individual differences in stability and short term change are related to both concurrent and future child antisocial behavior. MMSA is now available in commercial software but using MMSA presents challenges due to the computational intensity of the approach. Throughout the paper, we tried to give practical advice on applying MMSA in strategic ways to avoid the situation where the modeling procedure breaks down due to the computational burden.

Substantive Considerations

 Change in mean hazards session 1 to 2. Regardless of whether we focus on CSL estimates or latent log hazards from MMSA, 3 of the 6 hazards changed significantly from session 1 to 2. The pattern of change suggested that children became angrier and parents became more neutral and less positive from session 1 to 2. It appears that the second session was less enjoyable to both parents and children than the first. Further support for this possibility is the fact that the latent log hazard of parent negative behavior from the first session significantly and strongly predicted skipping the second session. It seems more likely that the parent rather than the child decided whether both would participate in session 2 or skip it and the parent’s decision was probably influenced by how much they enjoyed the first session. Parents with higher hazard rates of negativity probably found the first session more irritating and less enjoyable compared to parents with lower hazard rates, which contributed to their decision to skip the second.

 Retest reliabilities session 1 and 2. They started at .65 and went down as low as .16 for parent or child neutral. The retest reliability in our sample reflects 3 major sources of error, 1) variability from finite time sampling, 2) variability due to coder disagreement and 3) variability in true change due to time specific causes, all relative to true, time stable individual differences. Retest reliability is the bottom line for judging whether the reliability of a rate from a single session is adequate to proceed with standard regression modeling and our results indicate that it is not. It is also the key parameter necessary to judge whether combining the two sessions together to obtain a single rate is justified and our results indicate that it is not. In both cases, effects from standard regression models will be badly attenuated by the lack of reliability and in the case of combining both sessions, we run the additional risk of completely missing the effects of change. The remedy is the same in both cases and is straightforward, at least conceptually and consists of using a latent variable approach like MMSA to actually model the retest correlation and change and correct regression effects for attenuation due to the imperfect reliability.

 Often, published research reports give coder agreement as the only index of reliability. If coder agreement is adequate, then variables derived from the observation session are assumed to have adequate reliability. This is clearly insufficient information to make that judgment. Without knowing the retest reliability, there is no way to know if the observational variables have adequate reliability. Even if coders had perfect agreement, observational variables could still have inadequate retest reliability due to a combination of low FTS reliability and high time specific variation relative to time stable variation. There is just no way of knowing this from coder agreement alone.

 Correlations with antisocial variables. Despite the theoretical simplicity of the approach, 3 log hazard rates (parent negative, parent positive and child positive) of the 8 considered had significant correlations with fall K antisocial and 3 of the 8 had significant correlations with 3rd grade antisocial (parent negative, parent positive and change from session 1 to 2 in child neutral). The significant correlations for 2 of the 3 log hazards with 3rd grade antisocial held even after controlling for fall K antisocial (parent positive and change in child neutral). All of these significant relations were also apparent in MMSA models that were limited to a single hazard rate at both sessions. Although estimating one MMSA model with all 8 rates at both sessions and both antisocial variables would have been desirable, it was not computationally feasible.

 Simultaneous regression results. No clear pattern emerged from using all 8 of the CSL estimates of the log hazard rates as simultaneous predictors of 3rd grade antisocial controlling for fall K antisocial. In fact, none of the predictors were significant. This finding held even if an MMSA was used for the log hazard with the lowest FTS reliability, child anger instead of the CSL estimates. The masking of significant effects in the simultaneous regression equation suggests that looking for unique effects among the individual child and parent log hazard rates may be overly simplistic. Theoretically specified linear or nonlinear functions of individual log hazard rates might be more important. For example, the difference and sum of two hazards on the log scale corresponds respectively to the ratio and product of 2 hazards on the rate per unit of time at risk scale. Alternatively, using conditional hazard rates in dyadic social interaction may be more fruitful because the conditional rates incorporate information about both members of the dyad simultaneously (e.g., the hazard of child positive given parent positive). The downside of this, however, is that the number of potential rates increases dramatically which will also lead to reductions in retest reliability as the total number of transitions (e.g., child positive) get distributed across the different possible antecedent states (e.g., parent negative, neutral or positive).

Methodological Considerations

 The first and most important principle is to plan to get retest data (i.e., a second session) at least on a random subsample of the participants. Even with an hour of observation, high levels of FTS reliability and well trained coders, our retest reliabilities were well below standard recommended levels for use in regression models which leads to loss of power and attenuation of effect size. All dyadic social behavior should be assumed to be highly dependent on the immediate social contextual created on the spot by the dyad and hence have inadequate retest reliability even in highly structured problem solving paradigms until proven otherwise. Do not accept coder agreement as a substitute for retest reliability. Cost effective and statistically efficient designs for the observation of social behavior is an area in which more research is badly needed.

 Second, compute CSL estimates of log hazards and use small MMSA models to determine their reliability as proxies for latent log hazards. If the reliabilities are high, like .90 or higher, use them instead of creating latent log hazards with MMSA. This will vastly speed up computations and allow fitting more complex and realistic models. If the reliabilities are low, then CSL estimates might still work, especially if multiple sessions of observation are available but the CSL estimates will probably be badly censored (i.e., many zero rates). Models for censored normal variables are available in some SEM programs (e.g., Mplus) and if the censoring point is essentially the same for all subjects, these may be faster computationally than MMSA (see Stoolmiller, Eddy & Reid, 2000 for an example). This is another area that needs more research.

 MMSA might also benefit from newer, 2-step approaches to regression calibration that have appeared recently. For example, Skrondal and Kuha, (2012) focused on speeding up models for logistic regression when covariates are measured with error but the computational challenges due to high dimensional numerical integration are similar to MMSA.

 Finally, numerical integration is a procedure that can be accelerated by parallel processing, that is, using many processors or processing cores either in one compute or in a network of computers. This so called cluster computing is being rapidly developed for many different types of computationally intensive applications (e.g., Jacobs & Dinsmore, 2013) but evidently, psychologists have not yet felt the need. For important research questions, users of MMSA may need to spend the time and energy learning how to access and use these resources. In some ways, it should hardly come as a surprise that something as complex as human social interaction, even among dyads requires some of the most advanced computing capabilities to answer basic questions.

References

Dagne, G.A., & Snyder, J. (2009). Bayesian hierarchical duration model for repeated events: An application to behavioral observations. *Journal of Applied Statistics, 36*, 1267-1279.

Dagne, G.A., & Snyder, J. (2011). Relationship of maternal negative moods to child emotion regulation during family interaction. *Development and Psychopathology, 23*, 211-223.

Dishion, T.J., Forgatch, M., Van Ryzin, M., Winter, C. (2012). The Nonlinear Dynamics of Family Problem Solving in Adolescence: The Predictive Validity of a Peaceful Resolution Attractor. *Nonlinear Dynamics, Psychology, and Life Sciences, 16*, 331-352.

Gardner, W. (1993). Hierarchical continuous-time sequential analysis: A strategy for clinical research. *Journal of Consulting and Clinical Psychology*, *61*, 975-983.

Hollenstein, T., Granic, I., Stoolmiller, M., & Snyder, J. (2004). Rigidity in Parent–Child Interactions and the Development of Externalizing and Internalizing Behavior in Early Childhood. *Journal of Abnormal Child Psychology, 32*, 595–607.

Jacobs, W. & Dinsmore, T.W. (2013). Driving value from big data. Mountain View, CA: Revolution Analytics, Inc.

Lunkenheimer, E.S., Olson, S.L., Hollenstein, T.L., Sameroff, A.J., & Winter, C. (2011). Dyadic flexibility and positive affect in parent–child coregulation and the development of child behavior problems. *Development and Psychopathology 23*, 577–591.

Muthén, L.K. & Muthén, B.O. (2013). Mplus User’s Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Raudenbush, S. & Bryk, A. (2002). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.

Rogosa, D. & Ghandour, G. (1991). Statistical Models for Behavioral Observations. *Journal of Educational Statistics, 16*, 157-252.

Skrondal, A., & Kuha, J. (2012). Improved regression calibration. *Psychometrika, 77*, 649-669

Snyder, J., Stoolmiller, M., Wilson, M., & Yamamoto, M. (2003). Child anger regulation, parental responses to children’s anger displays, and early child antisocial behavior. *Social Development, 12*, 335–360.

Stoolmiller, M., Eddy, J.M., & Reid, J.B. (2000). Detecting and describing preventive intervention effects in a universal school-based randomized trial targeting delinquent and violent behavior. *Journal of Consulting and Clinical Psychology, 68*, 296-306.

Stoolmiller, M., & Snyder, J. (2006). Modeling Heterogeneity in Social Interaction Processes: Using Multilevel Survival Analysis. *Psychological Methods, 11*, 164–177.

Stoolmiller, M., & Snyder, J. (in press). Embedding Multivariate Multilevel Survival Analysis of Dyadic Social Interaction in Structural Equation Models: Hazard Rates as Both Outcomes and Predictors. *Journal of Pediatric Psychology.*

Tuma, N.B., & Hannan, M.T. (1984). Social dynamics: Models and Methods. New York: Academic Press, Inc.

Tables









Figure Captions

Figure 1. Path diagram for a simple MMSA with 2 log hazard rates and fall K child antisocial as a predictor and 3rd grade child antisocial as a distal outcome.

Figure 2. Path diagram for MMSA reliability (random intercept) model for 2 log hazard rates.

Figure 3. Path diagram for MMSA change (latent difference score) model for 2 log hazard rates.

Within Dyad Level

Between Dyad Level

Grade 3 Child Antisocial

Fall K Child Antisocial

Duration S2

NotAc→Ac

Censored? (yes=1,no=0)

 log f1Ac

Child Anger Session 1

Duration S1

NotAc→Ac

Censored?

(yes=1,no=0)

log f2Ac

Child Anger Session 2

Child Anger log residual session 1

log h0,Ac(t)

Child Anger

Base Hazard

Child Anger

log residual session 2

log h0,Ac(t)

Child Anger

Base Hazard

Within Dyad Level

Between Dyad Level

Duration S2

NotAc→Ac

Censored? (yes=1,no=0)

 log f1Ac

Child Anger Session 1

Duration S1

NotAc→Ac

Censored?

(yes=1,no=0)

log f2Ac

Child Anger Session 2

Child Anger log residual session 1

log h0,Ac(t)

Child Anger

Base Hazard

Child Anger

log residual session 2

log h0,Ac(t)

Child Anger

Base Hazard

 log fAc

Child Anger

log

Child Anger Intercept

log

Child Anger Difference

1

1

1

Child Anger log residual session 1

Within Dyad Level

Between Dyad Level

Duration S2

NotAc→Ac

Censored? (yes=1,no=0)

 log f1Ac

Child Anger Session 1

Duration S1

NotAc→Ac

Censored?

(yes=1,no=0)

log f2Ac

Child Anger Session 2

log h0,Ac(t)

Child Anger

Base Hazard

Child Anger

log residual session 2

log h0,Ac(t)

Child Anger

Base Hazard